Saarland University

Department of Computer Science

Studying User Experience and Acceptance of Web Authentication Solutions

Dissertation
zur Erlangung des Grades
des Doktors der Ingenieurwissenschaften
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

von
Sanam Ghorbani Lyastani

Saarbrücken, 2023

Tag des Kolloquiums:	05.02.2024
Dekan:	Prof. Dr. Jürgen Steimle

Prüfungsausschuss:

Vorsitzender:	Prof. Thorsten Herfet
Berichterstattende:	Prof. Dr. Dr. h.c. Michael Backes
	Prof. Bernd Finkbeiner
	Dr. Katharina Krombholz
Akademischer Mitarbeiter:	Dr. Zhiqiu Jiang

Abstract

To improve the security of their web authentication, users can employ password managers, set up two-factor authentication, or replace passwords with FIDO2 authenticator devices. However, for those solutions to be accepted by the user, their user experience must match the users' mental models.

This thesis contributes the novel methodologies and results of three studies that measured the user experience and acceptance of three web authentication solutions. Our results show that a) whether password managers are beneficial for security or aggravate existing problems depends on the users' strategies and how well the manager supports the users' individual password management right from the time of password creation; b) users consider FIDO2 passwordless authentication as more usable and more acceptable than password-based authentication, but also that impeding concerns remain that are rooted in a gap between the user's personal perspective onto this new technology and the global view of the FIDO2 designers; c) there is a lack of consistency between the two-factor authentication user journeys of top websites and that the more consistent design patterns are problematic for usability, which could increase users' cognitive friction and lead to rejection. Based on those results, we make suggestions for further research into understanding and improving the users' experience of web authentication.

Zusammenfassung

Um die Sicherheit ihrer Web-Authentifizierung zu verbessern, können Nutzer Passwortmanager einsetzen, eine Zwei-Faktor-Authentifizierung einrichten oder Passwörter durch FIDO2-Authentifizierung ersetzen. Damit diese Lösungen von den Nutzern akzeptiert werden, muss die Benutzererfahrung jedoch mit den mentalen Modellen der Nutzer übereinstimmen.

In dieser Dissertation stellen wir Methoden und Ergebnisse von drei Studien vor, in denen die Benutzererfahrung und die Akzeptanz von Web-Authentifizierungslösungen gemessen wurden. Unsere Ergebnisse zeigen, dass a) Passwortmanager die Sicherheit erhöhen oder bestehende Probleme verschärfen, abhängig von den Strategien der Nutzer und wie gut der Manager die individuelle Passwortverwaltung der Nutzer bereits bei der Passworterstellung unterstützt; b) Benutzer passwortlose FIDO2-Authentifizierung für benutzerfreundlicher und akzeptabler als Passwörter halten, aber Bedenken bleiben, die auf eine Diskrepanz zwischen der persönlichen Benutzerperspektive und der der FIDO2-Designer auf diese neue Technologie zurückzuführen sind; c) es bei der Zwei-Faktor-Authentifizierung Benutzererfahrung auf Top-Websites an Konsistenz fehlt, und die konsistenteren Designmuster problematisch für Benutzerfreundlichkeit sind, was die kognitive Belastung der Benutzer erhöht und zu Ablehnung führen könnte. Basierend auf diesen Ergebnissen machen wir Vorschläge für weitere Forschung zum Verständnis und zur Verbesserung der Nutzererfahrung bei Web-Authentifizierung.

Background of this Dissertation

This dissertation is based on the papers mentioned in the following. I contributed to all the papers as the main author.

The initial idea for the study in the first work [P1] was proposed by Sascha Fahl but was extended, implemented, and carried out by the author. Sven Bugiel helped with the implementation of the browser plugin and back-end server. Michael Schilling helped as empirical research support with the evaluation and statistical testing of our collected data. Both Michael Schilling and Sven Bugiel were involved in general writing tasks. Michael Backes provided feedback on the general direction of the project. All authors performed reviews of the paper.

The basic idea for the second work [P2] to study the acceptance of FIDO2 passwordless authentication was proposed by Sven Bugiel. The author developed the concrete concept and design for the user study and was responsible for executing the study. Michaela Neumayr assisted in the execution and helped with the coding of qualitative data. Michael Schilling helped as empirical research support with the evaluation and statistical testing of our collected data and provided feedback about the study design. Michael Schilling and Sven Bugiel were further involved in the process of writing the paper. All authors performed reviews of the paper.

The author proposed and developed the idea for the third work [P3] to study the consistency of two-factor authentication user journeys, and also executed the data collection and evaluation. Besides general discussions, Sven Bugiel was further involved in writing tasks and assisted with the evaluation. All authors performed reviews of the paper.

Author's Papers for this Thesis

[P1] Ghorbani Lyastani, S., Schilling, M., Fahl, S., Bugiel, S., and Backes, M. Studying the impact of managers on password strength and reuse. In: *Proc. 26th USENIX Security Symposium (SEC' 18)*. USENIX Association, 2018.

[P2] Ghorbani Lyastani, S., Schilling, M., Neumayr, M., Backes, M., and Bugiel, S. Is FIDO2 the kingslayer of user authentication? A comparative usability study of FIDO2 passwordless authentication. In: *Proc. 41st IEEE Symposium on Security and Privacy (SP '20)*. IEEE, 2020.

[P3] Ghorbani Lyastani, S., Backes, M., and Bugiel, S. A systematic study of the consistency of two-factor authentication user journeys on top-ranked websites. In: *Proc. 30th Annual Network and Distributed System Security Symposium (NDSS '23)*. The Internet Society, 2023.

Further Publications of the Author

[S1] Ghorbani Lyastani, S., Acar, Y., Backes, M., and Fahl, S. Poster: Improving password memorability and strength using mangling rules. In: *Proc. 12th Symposium on Usable Privacy and Security (SOUPS'16)*. USENIX Association, 2016.

Acknowledgments

First and foremost, I extend my deepest gratitude to my supervisor, Prof. Michael Backes, for offering me my first research internship at CISPA in 2014, which led to the incredible offer to pursue a PhD position in the Information Security & Cryptography group. He allowed me to conduct research in my areas of interest, continually motivating me to engage in the work I am passionate about. I am incredibly fortunate to be one of his students.

I want to express my special gratitude to my co-authors—Sven Bugiel, Michael Schilling, Sascha Fahl, and Michaela Neumayr—who have been pillars of support and knowledge throughout my PhD journey. Your collaboration, insights, and guidance have not only contributed significantly to our projects but have also enriched my learning and growth in this field.

I'm deeply grateful to Bettina Balthasar, whose kindness and support have been unwavering since my early days in Germany. Notably, she was the first to warmly welcome me in English, breaking the silence of my initial two days in a new country. I'll always appreciate the warm smiles she gave whenever I appeared at her door's office. Thank you for adding so much joy and easing the stress along this journey.

Huge thanks to my awesome officemates—Duc, Michaela, and Michael— you guys totally rocked our workspace, turning it into a scene straight out of 'Friends' rather than the usual 'Office' vibes. Thanks for being the best colleagues-turned-friends I could ask for.

To the members of our group, thank you for the countless fantastic moments we've shared together. Additionally, my thanks go out to the staff across various departments at CISPA for their assistance with paperwork, IT matters, and business travels.

To my parents, thank you for your endless love, support, and belief in me. Your sacrifices and encouragement have been my constant source of strength and motivation. Thank you for everything you've done, for the seen and unseen ways you've helped me reach this point. My heartfelt thanks go to my sister, who has been an unwavering source of support, laughter, and wisdom throughout my journey. Her belief in me, even when I doubted myself, her endless encouragement, and her ability to make me laugh during the most stressful times have been invaluable. Her presence in my life is a gift I cherish deeply, and I am endlessly grateful for her love and companionship. Thank you for being not just a sister but a true friend. A special mention to my niece, Vina. Your arrival and growth during these crucial years have brought joy, laughter, and a much-needed perspective on life beyond work and study. Thank you for being my unexpected source of inspiration and for making this phase of my life infinitely richer.

Last but not least, thanks to my fiancé, Sven, for your understanding, patience, and endless love. Your support and belief in my capabilities have been my anchor during the most challenging times. You've kept me grounded, nourished, and focused, even when stress made me lose sight of the basics. I couldn't have navigated this path without you by my side. Thank you for being my constant source of love and motivation.

Contents

List of Figures

List of Tables

1
Introduction

Authentication is one of the most fundamental security goals for protecting online services. Only with reliable authentication can it be ensured that online services and their related data can solely be accessed by authorized persons. Despite the wide range of authentication options that exist today, passwords are the most widespread ones. According to history and literature, passwords have been around for centuries and can be traced back to ancient times and the biblical Book of Judges [156]. The Greek historian Polybius explained how Romans used a "watchword" to distinguish enemies from friends. In the 1960s, the idea of passwords for computer systems was introduced by Fernando J. Cobató [67, 216], an American pioneer in the field of time-sharing operating systems, who is also known as the father of modern passwords during his work at MIT. He presented the concept of passwords for the Compatible Time-Sharing System operating system to help keep individual files of users private on the shared computer system.

A text-based password is a shared secret character string used to verify the identity of a user during the authentication process. Initially, in the age of the ARPANET, passwords have been applied only in research and lab settings; the security flaws inherent to passwords were discovered when the computer gradually started to make its way into homes and offices, and ordinary users had to adopt password authentication. For instance, the Morris Worm [187], one of the oldest malware, could spread in 1988 by exploiting weak user passwords. By now, with the increasing digitalization of our society, there is an observable trend toward an increasing number of online services that users register to—all demanding a form of user authentication. This increasing number of required passwords, in combination with the limited human capacity to remember passwords, leads to the insecure practice of reusing easily guessable passwords across accounts [75, 154, 37, 196]. Fernado J. Cobató, the inventor of the modern password on computer systems, reflected at the age of 87 years in an interview with The Wall Street Journal [216] on passwords:

> *"Unfortunately it's [password] become kind of a nightmare with the World Wide Web. I don't think anybody can possibly remember all the passwords that are issued or set up. That leaves people with two choices. Either you maintain a crib sheet, a mild no-no, or you use some sort of program as a password manager. Either one is a nuisance."*

This problem has been underlined by research efforts [139] that, again and again, demonstrated that passwords are poor practice when it comes to security. As a result, the security community has tried to improve or replace purely text-based passwords with more secure alternatives for end-user authentication on the web [15]. Among the different authentication schemes brought forward to the password problem, there are generally three types of authentication: 1) something you know (*knowledge factor*), e.g., passwords, 2) something you have (*possession factor*), e.g., mobile phone, and 3) something you are (*inherence factor*), e.g., fingerprint. To remedy the drawbacks of text-based passwords, those solutions either a) provide technical help for users in the form of password managers to securely generate, store, and enter passwords; or b) they necessitate additional steps, such as two-factor authentication, so that the security of the text-based password alone is not decisive anymore; or c) they entirely replace text-based passwords with token-based authentication, such as the nowadays widely supported WebAuthn by the FIDO Alliance.

For these solutions to be successful in practice, they must be understood and accepted by the users. The users must have a good experience with the methods and be willing to use those solutions correctly. Otherwise, the security benefits of any solution remain ineffective.

In this thesis, we present a line of work that studies three solutions for the password problem from the users' perspective regarding experience, acceptance, and correct usage.

Summary of contributions

Chapter 2: Better managed than memorized? Despite their well-known security problems, passwords are still the incumbent authentication method for virtually all online services. To remedy the situation, users are very often referred to password managers as a solution to the password reuse and weakness problems. However, to date, the actual impact of password managers on password strength and reuse has not been studied systematically. In our USENIX Security'18 work [P1], we provided the first large-scale study of password managers' influence on users' real-life passwords. By combining qualitative data on users' password creation and management strategies, collected from 476 participants of an online survey, with quantitative data (incl. password metrics and entry methods) collected in situ with a browser plugin from 170 users, we were able to gain a more complete picture of the factors that influence our participants' password strength and reuse. Our approach allowed us to quantify for the first time that password managers indeed influence password security. However, whether this influence is beneficial or aggravates existing problems depends on the users' strategies and how well the manager supports the users' password management right from the time of password creation. Given our results, we think research should further investigate how managers can better support users' password strategies in order to improve password security as well as stop aggravating the existing problems.

Chapter 3: Is FIDO2 the kingslayer of user authentication? The newest contender for succeeding passwords as the incumbent web authentication scheme is the FIDO2 standard. Jointly developed and backed by the FIDO Alliance and the W3C, FIDO2 has found support in virtually every browser, finds increasing support from service providers, and has adoptions beyond browser software on its way. While it supports multi-factor and two-factor authentication, its single-factor, passwordless authentication with security tokens has received the bulk of attention and was hailed by its supporters and the media as the solution that will replace text-based passwords on the web. Despite its obvious security and deployability benefits—a setting that no prior solution had in this strong combination—the paradigm shift from a familiar knowledge factor to purely a possession factor raises questions about the acceptance of passwordless authentication by end-users.

Our IEEE Security & Privacy'20 work [P2] presented the first large-scale lab study of FIDO2 single-factor authentication to collect insights about end-users perception, acceptance, and concerns about passwordless authentication. Through hands-on tasks, our participants gather first-hand experience with passwordless authentication using a security key, which they afterward reflect on in a survey. Our results showed that users are willing to accept a direct replacement of text-based passwords with a security key for single-factor authentication. That is an encouraging result in the quest to replace passwords. But, our results also identified new concerns that can potentially hinder the widespread adoption of FIDO2 passwordless authentication. In order to mitigate these factors, we derived concrete recommendations to try to help in the ongoing proliferation of passwordless authentication on the web.

Chapter 4: Are 2FA user journeys on top-ranked websites consistent? Heuristics for user experience state that users will transfer their expectations from one product to another. A lack of consistency between products can increase users' cognitive friction, leading to frustration and rejection. This work from NDSS'23 [P3] presented the first systematic study of the external, functional consistency of two-factor authentication user journeys on top-ranked websites. We found that these websites implement only a minimal number of design aspects consistently (e.g., naming and location of settings) but exhibit mixed design patterns for setup and usage of a second factor. Moreover, we found that some of the more consistently realized aspects, such as descriptions of two-factor authentication, have been described in the literature as problematic and adverse to user experience. Our results advocate for more general UX guidelines for 2FA implementers and raise new research questions about the 2FA user journeys.

Outline

The remainder of this dissertation is structured as follows. We present our study of password manager usage in Chapter 2 and our study of FIDO2 passwordless authentication in Chapter 3. Our study of the consistency of two-factor authentication user journeys is presented in Chapter 4. We conclude this dissertation in Chapter 5.

2
Password Manager Study

Better managed than memorized? Studying the Impact of Managers on Password Strength and Reuse

2.1 Motivation

For several decades passwords prevail as the default authentication scheme for virtually all online services [131, 15, 91]. At the same time, research has again and again demonstrated that passwords perform extremely poor in terms of security [139]. For instance, various attacks exploit that humans fail to create strong passwords themselves [14, 55, 133, 92, 96]. Even worse, there is an observable trend towards an increasing number of online services that users register to. This increasing number of required passwords in combination with the limited human capacity to remember passwords leads to the bad practice of re-using passwords across accounts [75, 154, 37, 196].

In the past, different solutions have been implemented to help users creating stronger passwords, such as password meters and policies, which are also still subject of active research [111, 172, 24, 133, 205]. Among the most often recommended solutions [85, 183, 168, 190, 178] to these problems for end-users is technical support in the form of password management software. Those password managers come built-in to our browsers, as a browser plugin, or as separate applications. Password managers are being recommended as a solution because they fulfill important usability and security aspects at the same time: They store all the users' passwords so the users do not have to memorize them; they can also help users entering their passwords by automatically filling them into log-in forms; and they can also offer help in creating unique, random passwords. By today, there are several examples of third party password managers that fit this description, such as Lastpass [121], 1Password [1], and even seemingly unrelated security software, such as anti-virus [106] solutions.

2.2 Problem Description

Unfortunately, it has not been sufficiently studied in the past whether password managers fulfill their promise and, indeed, positively influence password security. To break this question down, we are interested in *1) whether password managers actually store strong passwords that are likely auto-generated by, for instance, password generators, or if they really are just storage where users save their self-made, likely weak passwords?* Further, we are interested whether *2) users, despite using password managers, still reuse passwords across different websites or if do they use the managers' support to maintain a large set of unique passwords for every distinct service?* Prior works [196, 154] that studied password reuse and strength in situ have also considered password managers as factors, but did not find an influence by managers and could not conclusively answer those questions.

We argue that to specifically study the impact of password managers, important aspects were missing in prior work, and this work's most tangible contribution is an extension of prior methodologies to be able to study password managers' impact in the wild. First, previous works considered only the presence of password management software on the user device and whether a password was auto-filled or not. However, to better distinguish the storage option of a password (i.e., memorized and manually entered, auto-filled by the browser, copy&pasted, or filled by a browser plugin) a more fine-grained entry method detection is required. Second, users do not axiomatically follow strict workflows for password creation, storage, and entry [78, 89, 190, 178, 181] (see Figure 2.1). For instance, the effort users are willing to invest in creating a unique and strong password often depends on the privacy-sensitivity of the associated account. For creating a new password, the approaches range from mental algorithms (e.g., leetifying a known word) over pen&paper algorithms and password generator tools (e.g., websites like `https://www.random.org/passwords/`) to 3rd party password managers (e.g., LastPass, KeePass, etc.). Based on different factors, such as technical skills, trust in software vendors, financial expenditure, multi-device support, or others, users resort to different password storage options from where the password finds its way via various entry methods into the login forms. To better study password managers' influence, one has to take the users' creation and storage strategies into consideration as well. In particular, one has to understand if the user pursues primarily a creation strategy based on password manager support and whether there then exists an observable effect of this strategy on the password strength and reuse.

2.3 Contributions

In this work, we present a study that reflects those considerations (see the bottom of Figure 2.1). We first recruited 476 participants on Amazon MTurk to conduct a survey sampling to better understand users' strategies for creating and storing passwords, their attitudes towards passwords, and past experiences with password leaks or password managers. From those insights, we identified two distinct groups in our participant pool: users of password managers and users abstaining from technical help in password creation. We were further able to recruit 170 of our participants, 49 of which reported using password managers, for a follow-up study in which our participants allowed us to monitor their passwords through a Google Chrome browser plugin that collected password metrics as well as answers to in situ questionnaires upon password entry. This gave us detailed information about real-life passwords, including their strength, their reuse, and, for the first time, their entry method (e.g., manually typed, auto-filled, pasted, or entered by a browser plugin) as well as the passwords' context, including user reported value of the password (e.g., loss of social repudiation or financial harm when the password would be leaked).

Based on the combined data from our survey sampling and plugin-based data collection, we are able to study the factors that influence password strength and reuse from a new perspective. Using exploratory data analysis and statistical testing, including regression models, we are first to actually show that password managers indeed influence password strength and reuse. In particular, the relation between different entry methods and the password strength depends on the users' entire process of password handling. Using a workflow that includes technical support from password creation through storage to entry leads to stronger passwords, while this positive effect on password strength cannot be detected when considering the input method individually. A similar picture emerges for password reuse. Passwords entered manually or by Chrome auto-fill were unique in only 20–25% of all cases. For LastPass or Copy&Paste password entry, the proportion of non-reused passwords increases to 53–78%. This is still far from ideal—that not even a single password is reused—but still a significant improvement

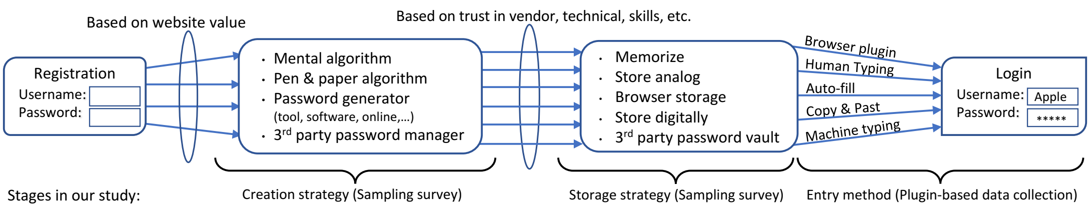


Figure 2.1: Users' strategies for password creation and storage plus the stages of our study to investigate managers' influence.

through such dedicated password management tools. Similar to the results for password strength, we find that password reuse improves further if the password generation is technically supported. In contrast to password strength, however, this positive effect is similar for all input methods. Looking at managers that do not offer support for password creation, such as Chrome's auto-fill, we even found a negative influence in that those managers even contribute to the password reuse problem. In summary, our results support the fact that technical tools can have a very positive effect on password security. However, it is important that the entire password management process is supported—from generation, over storage, to entry—and not only the old and weak passwords of the users are stored.

2.4 Methodology

For our study of password managers' impact on password strength and reuse, we use data collected from paid workers of Amazon's crowd-sourcing service *Mechanical Turk*. We collected the data in two different stages: 1) a survey sampling, and 2) collection of *in situ* password metrics.

Ethical concerns: The protocols implemented in those two stages were approved by the ethical review board[1] of our university. Further, we followed the guidelines for academic requesters outlined by MTurk workers [57]. All server-side software (i.e., a LimeSurvey installation and a self-written server application) was self-hosted on a maintained and hardened university server. Web access to the server was secured with an SSL certificate issued by the university's computing center and all further access was restricted to the department's intranet and only made available to maintainers and collaborating researchers. Participants could leave the study at any time.

2.4.1 Password survey

In our survey sampling, we asked participants about their general privacy attitude, their attitude towards passwords, their skills and strategies for creating and managing passwords, as well as basic demographic questions. Those information enable us, on the one hand, to gain a general overview of common password creation and storage strategies. On the other hand, those information help us in detecting and avoiding

[1] https://erb.cs.uni-saarland.de/

any potential biases in the later stages of our study. The full survey contained 31–34 questions, categorized in 6 different groups (see Appendix A.1).

We first asked for their privacy attitude using the standard Westin index [116]. However, since the Westin index has been shown to be an unreliable measure of the actual privacy-related actions of users [211], we also asked about the participants' attitude towards passwords (e.g., whether they consider passwords to be futile in protecting their privacy).[2] This should help in better understanding if participants are actually motivated to put an effort into creating stronger and unique passwords. We further asked about the participants' strategies for password creation and management in order to get a more complete picture of the possible origins of passwords in our dataset.

All qualitative answers (e.g., *Q9* or *Q22* in Appendix A.1) were independently coded in a bottom-up fashion by two researchers. The researchers achieved an initial agreement between 95.6% (*Q9*) and 97.1% (*Q22*) and all differences could be resolved in agreement.

Participation in the survey was open to any MTurk worker that fulfilled the following criteria: the worker was located in the US and the number of previously approved tasks was at least 100 or at least 70% all of the tasks. The estimated time for answering the survey was 10–15 minutes and we paid $4 for participation.

In total, 505 MTurk workers participated in our survey between August 2017 and October 2017. After discarding responses that failed attention test questions [95], were answered too fast to be done thoughtfully, or that were duplicates, we ended up with 476 valid responses.

Lastly, we also asked whether the participant would be willing to participate in a follow-up study, in which we measure in an anonymized, privacy-protecting fashion the strength and reuse of their passwords. Only participants that indicated interest in the follow-up study were considered potential candidates for our Chrome plugin-based data collection. Only 21 workers were not interested.

2.4.2 Chrome plugin-based data collection

To collect in situ data about passwords, including strength, reuse, entry method, and domain, we created a Chrome browser plugin that monitors the input to password fields of loaded websites and then sends all collected metrics back to our server once the user logs in to the website. We distributed our plugin via the Google Web Store to invited participants. The plugin was unlisted in the Store, so that only participants to which we sent the link to the plugin store website were able to install it. Our primary selection criterion for participant selection was that they use Chrome as their primary browser and are not using exclusively mobile devices (smartphones and tablets) to browse the web; besides that we aimed for an unbiased sampling from the participants pool with respect to the participants' privacy attitude, attitude towards passwords, demographics, and usage of password managers. Between September and October 2017, we invited

[2] Other instruments, which meet the latest requirements of scale constructions and which are often used in recent research, do not reflect the actual privacy/security attitude construct, but refer more strongly to security behavior (e.g., SeBIS [60]) or are strongly tailored to the corporate context (e.g., HAIS-Q [152]).

364 participants from the survey sampling to the study, of which 174 started and 170 finished participation. We asked participants to keep our plugin installed for at least four days. Participants that finished the task were compensated with $20. Our plugin collects the following metrics:

Composition: The length of the entered password as well as the frequency of each character class.

Strength: The password strength measured in Shannon and NIST entropy as well as zxcvbn score. Shannon and NIST entropy have been used in prior works [64, 196, 61] as a measure of password strength and complexity and are collected primarily to be backward compatible in our analysis with prior research. However, since entropy has been shown to be a poor measurement of the actual "crackability" of the password [203], we use the zxcvbn [205] score as the more realistic estimator of the password strength in our analysis.[3] Zxcvbn estimates every password's strength on a scale from 0 (weakest) to 4 (strongest) using pattern matching (e.g., repeats, sequences, keyboard patterns), common password dictionaries (including leaked passwords, names, English dictionary words), and mangling rules (e.g., leetify). Appendix A.2 explains the meaning of this score in more detail. In our plugin we used the zxcvbn library [54] with its default settings. From a statistical point of view, a metrically scaled strength measurement instead of the ordinal zxcvbn score would have helped in finding possible effects on password strength easier (see Section 2.5), however, it does not affect the presence of possible effects per se.

Website category: The category of the website domain according to the *Alexa Web Information Service* [7]. Our plugin contains the category for the top 28,651 web domains at the time the study was conducted.[4]

Entry method: The method through which the password was entered, such as *human*, *Chrome auto-fill*, *copy&paste*, *3rd party password manager plugin*, or *external password manager program*. The detection of the entry method is described separately in Section 2.4.2.1.

In situ questionnaire: Participant's answers to a short questionnaire about the entered password and website (see Section 2.4.2.2). In particular, we ask about the website's *value* for their privacy. Other studies used the website category as a proxy for this value [154] and in our study we wanted first-hand knowledge (see also Appendix A.3).

Hashes: Adapting the methodology of [154, 196], we collect the hash of the entered password as well as the hash of every 4-character sub-string of the password. We use a keyed hash (i.e., PBKDF2 with SHA-256), where the key is generated and stored at the client side and never revealed to us. This allows identification of (partially) reused passwords *per participant.* We use the notions introduced in [154]: *Exactly reused* passwords are identical with another password, *partially reused* passwords share a sub-string with another password, and *partially-and-exactly reused* passwords have both of those characteristics. Like related work [154, 196], we cannot compare passwords across participants.

[3] Unfortunately, the fully trained neural network based strength estimator of [154, 133] was not publicly available.

[4] This is the number of web domains in the top 100K list, for which a category was assigned by Alexa.

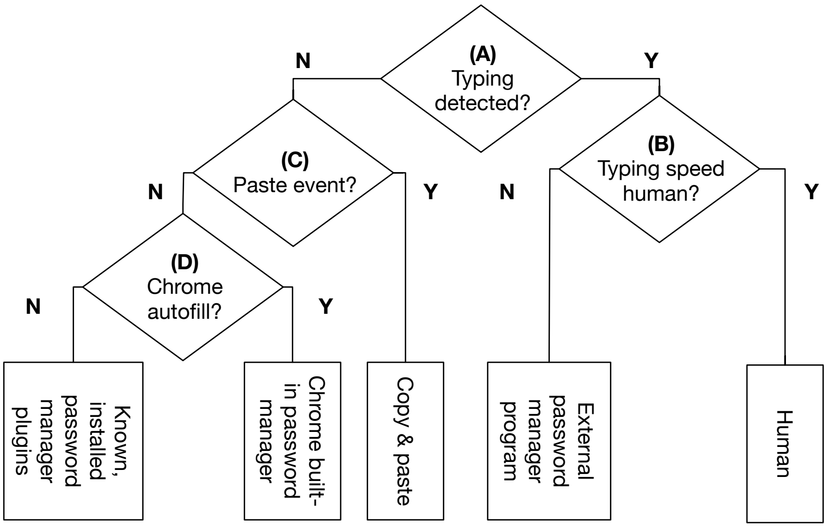

(A)
Typing detected?
N
Y
(C)
Paste event?
N
Y
(B)
Typing speed human?
N
Y
(D)
Chrome autofill?
N
Y
Known, installed password manager plugins
Chrome built-in password manager
Copy & paste
External password manager program
Human

Figure 2.2: Decision tree to detect password entry methods.

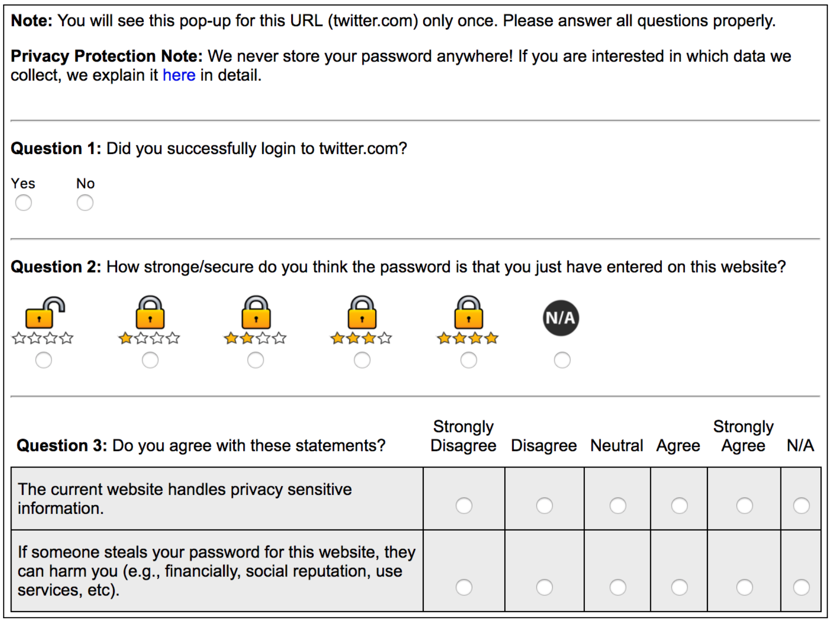
2.4.2.1 Detecting the entry method

Detection of the password entry method follows the decision tree depicted in Figure 2.2. If our plugin detects any kind of typing inside the password field (**(A)=Y**) and the typing speed is too fast to come from a human typist (**(B)=N**), we conclude that an external password manager program (such as KeePass) mimics a human typist by "replaying" the keyboard inputs of the password. Otherwise (**(B)=Y**), we assume a manually entered password. As threshold between human and external program, we set an average key press time of 30 ms. This is based on the observation that external programs usually do not consider mimicking the key press time, while some of them enter the password character-wise with varying speeds. In case there was no typing detected (**(A)=N**) and a paste event was observed (**(C)=Y**), we consider the password to be pasted by either a human or an external program. In either case, the password is managed externally to the browser in digital form. If no paste event was detected (**(C)=N**) and the Chrome auto-fill event was observed, this indicates that Chrome filled the password field from its built-in password manager. If Chrome auto-fill has not filled the password field (**(D)=N**), our plugin checks the list of installed plugins for eight well-known password manager plugins (see Appendix A.4) and reports the ones installed in the participant's browser, or an "unknown" value in case none of those eight was found.

We make the assumption that the user does not enter the password with a mixture of the different entry methods (e.g., pasting a word and complementing it with typing). Such mixture of entry methods would result in misclassification of the detected method. However, we assume that such behavior is too rare to affect our results significantly.

2.4.2.2 Participant instructions

We provided our participants with a project website that gave a step-by-step introduction on how to install our plugin, set it up, use it, and remove it post-participation. Google



Note: You will see this pop-up for this URL (twitter.com) only once. Please answer all questions properly.

Privacy Protection Note: We never store your password anywhere! If you are interested in which data we collect, we explain it here in detail.

Question 1: Did you successfully login to twitter.com?

Yes	No
○	○

Question 2: How stronge/secure do you think the password is that you just have entered on this website?

☆☆☆☆	★☆☆☆	★★☆☆	★★★☆	★★★★	N/A
○	○	○	○	○	○

Question 3: Do you agree with these statements?	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	N/A
The current website handles privacy sensitive information.	○	○	○	○	○	○
If someone steals your password for this website, they can harm you (e.g., financially, social reputation, use services, etc).	○	○	○	○	○	○

Figure 2.3: In situ questionnaire upon login to a new website.

Web Store provided our participants with a very comfortable way of adding the plugin to their browser. To set the plugin up, participants had to simply enter their MTurk worker ID into the plugin. The worker ID was used as a pseudonym throughout this study to identify data of the same participant. After setup, the plugin starts monitoring the users' password entries. For every newly detected domain to which a password was submitted, our plugin asked the participant to answer a short three question questionnaire about the participants' estimate of the website's value, the participants' strength estimate of the just entered password, and whether the login was successful (see Figure 2.3). Every participant was instructed to use the plugin for four days, after which the plugin released a completion code to be entered into the task on MTurk to finish participation and collect the payment. Through our server logs and the Google Web Store Developer Dashboard we confirmed that all participants removed our plugin shortly after finishing participation. We also instructed participants to act naturally and not change their usual behavior during those four days in order to maximize the ecological validity of our study. The only exceptions from the usual behavior were the installation of our plugin and a request to re-login to all websites where they have an account in order to ensure a sufficient enough quantity of collected data.

2.4.2.3 Addressing privacy concerns

A particular consideration of our study design was the potential privacy concerns of our participants. Since we essentially ask our participants to install a key-logger that monitors some of the most privacy-sensitive data, this might repel participants from participating. Due to the lack of in-person interviews or consultation between the researchers and the participants, we tried to address those concerns through a high level of transparency, support, and collecting only the minimal amount of data in a privacy-protecting fashion, which also follows the guidelines for academic requesters [57].

First, we explained on our project website the motivation behind our study and why acting naturally is important for our results. In this context, we provided a complete list of all data that our plugin collects, for which purpose, and why this data collection does not enable us to steal the participants' passwords. We also answered all participants' questions in this regard that were sent to us via email or posted in known MTurk review/discussion forums. We received feedback from workers that this level of openness has convinced them to participate in the study. Second, we distributed our plugin in an authenticated way via the Google Web Store and did not obfuscate the plugin's code. Third, we limited the extent of the collected data to the necessary minimum while still being able to study password managers' impact. For instance, we only collect the first successful login to any website, thus abstaining from monitoring participants' browsing behavior. Fourth, every participant could inspect the collected data per domain prior to sending them to us and chose to skip the data collection for highly sensitive websites.

2.5 Studying Password Managers' Impact

In this section, we analyze our collected data, but leave the discussion of our results for Section 2.6. After presenting our participants' demographics and an overview of their password reuse and strength, we group our participants based on their creation strategy and study the impact of different password management and creation strategies.

2.5.1 Demographics

Table 2.1 provides an overview of the demographics of our participants that answered our survey, that we invited to the plugin-based study, and that participated in the plugin-based data collection. We invited participants in equal parts from every demographic group and every demographic group participated in almost equal parts in the plugin-based data collection. We use a Mann-Whitney rank test [74] to test for significant differences between the demographic distributions of the 476 participants in the survey sampling and the 170 participants in the plugin-based study, and could not find any statistically significant ($p < .05$) differences between those two groups. In general, our participants' demographics are closer to the commonly observed demographics of qualitative studies in university settings than to the demographics of the 2010 US census [186]. Our participant number is skewed towards male participants (57.6% identified themselves as male). Also, our participants covered an age range from 18 to more than 70 years, where our sample skews to younger participants (75.2% of our study participants are younger than 40) as can be commonly observed in behavioral research, including password studies and usable security. The majority of our participants had no computer science background (80.88%) and was English speaking (98.3%). Most of the participants identified themselves as of white/Caucasian ethnicity (74.6%). The participants also covered a range of educational levels, where a Bachelor's degree was the most common degree (36.6% of all participants). Further, 80.9% of our participants reported using Chrome as their primary browser (see Table 2.2).

	Survey	**Invited to study**	**Participated**
Number of participants	476	364	170
Gender			
Female	200	156 (78.0%)	73 (36.5%)
Male	274	208 (75.9%)	97 (35.4%)
Other	1	0	0
No answer	1	0	0
Age group			
18–30	180	139 (77.2%)	64 (35.6%)
31–40	178	135 (75.8%)	63 (35.4%)
41–50	71	58 (81.7%)	32 (45.1%)
51–60	35	24 (68.6%)	8 (22.9%)
61–70	11	7 (63.6%)	2 (18.2%)
≥71	1	1 (100%)	1 (100%)
Computer science background			
Yes	91	64 (70.3%)	27 (29.7%)
No	385	300 (77.9%)	143 (37.1%)
Native language			
English	468	358 (76.5%)	167 (35.7%)
Other	8	6 (75.0%)	3 (37.5%)
Education level			
Less than high school	3	3 (100%)	1 (33.3%)
High school graduate	68	53 (77.9%)	26 (38.2%)
Some college, no degree	117	85 (72.6%)	34 (29.1%)
Associate's degree	79	64 (81.0%)	34 (43.0%)
Bachelor degree	174	133 (76.4%)	62 (35.6%)
Ph.D	2	1 (50.0%)	1 (50.0%)
Graduate/prof. degree	32	25 (78.1%)	12 (37.5%)
Other	1	0	0
Ethnicity			
White/Caucasian	355	274 (77.2%)	123 (34.6%)
Black/African American	50	38 (76.0%)	25 (50.0%)
Asian	31	23 (74.2%)	9 (29.0%)
Hispanic/Latino	27	21 (77.8%)	12 (44.4%)
Native American/Alaska	1	0	0
Multiracial	7	5 (71.4%)	1 (14.3%)
Other	5	3 (60.0%)	0

Table 2.1: Demographics of our participants. Percentages indicate the fraction w.r.t. initial size in the survey sampling.

Browser	Chrome	Firefox	Safari	Opera	IE/Edge	Other
Share	385 (80.9%)	71 (14.9%)	7 (1.5%)	6 (1.3%)	1 (0.2%)	6 (1.3%)

Table 2.2: Primary browsers of our 476 survey participants.

	Survey	Invited to study	Participated
Privacy concern (Westin index)			
Fanatic	217	167 (77.0%)	66 (30.4%)
Unconcerned	86	56 (65.1%)	31 (36.0%)
Pragmatist	173	141 (81.5%)	73 (42.2%)
Attitude about passwords			
Pessimist	9	8 (88.9%)	3 (33.3%)
Optimist	365	279 (76.4%)	132 (36.2%)
Conflicted	102	77 (75.5%)	35 (34.3%)
Prior password leak experienced			
No	190	151 (79.5%)	72 (37.9%)
Yes	148	111 (75.0%)	58 (39.2%)
Not aware of	138	102 (73.9%)	40 (29.0%)

Table 2.3: Privacy attitude, attitude about passwords, and prior experience with password leakage among our participants.

Since our study effectively asks participants to install a password-logger, we were concerned with a potential opt-in bias towards people that have low privacy concerns or consider passwords as ineffective security measures. To this end, we included the three questions of Westin's Privacy Segmentation Index [116] (*Q1* in Appendix A.1) to capture our participants' general privacy attitudes (i.e., fundamentalists, pragmatists, unconcerned). We further added two questions specifically about our participants' attitude about passwords (see *Q4* in Appendix A.1), e.g., if passwords are considered a futile protection mechanism or important for privacy protection. Table 2.3 summarizes the results of those questions. Only a minority of 86 of our survey participants are privacy unconcerned and the majority of 365 participants believe in the importance of passwords as a security measure. Almost a third of our survey participants experienced a password leak in the past. For our study we sampled in almost equal parts from those different groups. Using a Mann-Whitney rank test, we could not find any statistically significant differences between the survey and study participants' distribution of privacy and password attitudes/experiences. Thus, we argue that the risk of an opt-in bias towards either end of the spectrum for privacy and password attitude is unlikely.

2.5.2 General password statistics

Tables 2.4 and 2.5 provide summary statistics of all passwords collected by our plugin. We collected from our 170 participants 1,045 unique passwords and 1,767 password entries in total. That means, that our average participant entered passwords to 10.39 distinct domains with a standard deviation of 5.52 and median of 9. Our participants reported using on average 29.95 password-secured accounts (*Q2* in Appendix A.1) and

Statistic	**Mean**	**Median**	**SD**	**Min**	**Max**
No. of passwords	10.39	9.00	5.52	1.00	27.00
Entry methods	2.24	2.00	0.75	1.00	4.00
Percentage reused passwords					
Non-reused	29.44%	21.58%	28.25%	0.00%	100%
Only-exact	15.72%	0.00%	24.43%	0.00%	100%
Only-partially	18.38%	11.11%	19.88%	0.00%	100%
Exact-and-partial	36.46%	38.75%	30.88%	0.00%	100%
Password composition					
Length	9.61	9.29	1.72	6.33	16.86
Character classes	2.52	2.50	0.58	1.00	3.94
Digits	2.54	2.38	1.24	0.25	6.73
Uppercase letters	0.85	0.67	0.81	0.00	4.62
Lowercase letters	5.92	5.72	1.96	1.67	15.50
Special chars	0.30	0.10	0.54	0.00	5.19
Password strength					
Zxcvbn score	2.20	2.14	0.75	0.67	4.00
Shannon entropy	29.31	28.37	7.93	16.00	68.00
NIST entropy	23.50	23.00	2.98	17.17	35.69

Table 2.4: Summary statistics for all 170 participants in our plugin-based data collection. We first computed means for each participant and then computed the mean, median, standard deviation, and min/max values of those means.

we collected on average 61% of each participant's self-estimated number[5] of passwords. The lowest number of domains per participant is 1 and the highest is 27, where the 1^{st} quartile is 6 and the 3^{rd} quartile is 14. Those numbers are hence slightly lower than those reported in related studies [154]. When considering only unique passwords, our average participant has 6.15 passwords, indicating that passwords are reused frequently. Our participants entered their passwords on average with 2.24 different methods. Looking at all passwords, our participants reused on average 70.56% of their passwords, where exact-and-partial reuse is most common with 36.46% of all passwords. Interestingly the minimum and maximum in all reuse categories is 0% and 100%, respectively, meaning that we have participants that did not reuse any of their passwords as well as participants that reused all of their passwords. The average password in our dataset had a length of 9.61 and was composed of 2.52 character classes. The average zxcvbn score was 2.20, where the participant with the weakest passwords had an average of 0.67 and the participant with the strongest an average of 4.00. Like prior work [196], we observe a significant correlation between password strength and reuse (chi-square test: $\chi^2 = 75.48$, $p < .001$).

As shown in Table 2.5, the majority of the 1,767 logged passwords were entered with Chrome auto-fill (53.71%) followed by manual entry (33.39%). Although in our pilot study various password manager plugins, e.g., KeePass and 1Password, had been correctly detected, in our actual study only LastPass was used by our participants. Of all passwords, 128 (7.24%) were entered with LastPass, which is a similar share

[5] Some participants underestimated this number

Entry method	All passwords	Unique passwords
Chrome auto-fill	949 (53.71%)	540 (51.67%)
Human	590 (33.39%)	331 (31.67%)
LastPass plugin	128 (7.24%)	100 (9.57%)
Copy&paste	55 (3.11%)	51 (4.88%)
Unknown plugin	41 (2.32%)	23 (2.20%)
External manager	4 (0.23%)	0 (0.00%)
$\sum$	1,767	1,045

Table 2.5: Number of password entries with each entry method.

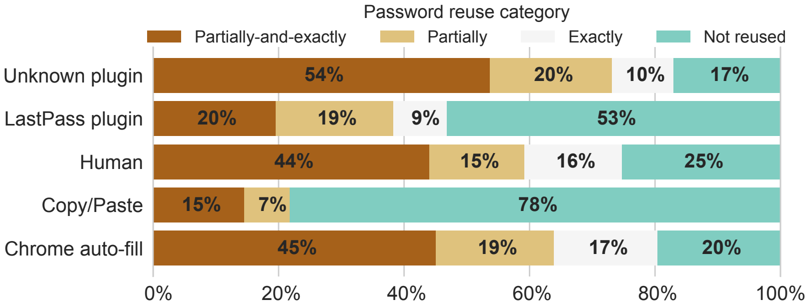


Figure 2.4: Password reuse by entry method for all passwords.

of managers as in recent reports [136]. Copy&paste and unknown plugins formed the smallest, relevant-sized shares and only four passwords were entered programmatically by an external program.

With respect to general password reuse (see Figure 2.4), partial-and-exact reuse is by far the most common reuse across all entry methods, except for LastPass' plugin and Copy&paste, which have a noticeably high fraction of non-reused passwords (e.g., 68 or 53% of all passwords entered with LastPass were not reused) and have noticeably less password reuse than the overall average. Looking at the password strength for all *unique* passwords (see Figure 2.5), one can see that 65% or 44 of all passwords entered with LastPass are stronger than the overall average of 2.20, while the other entry methods show a more balanced distribution across the zxcvbn scores (except for score 0). In summary, this indicates that LastPass shows an improved password strength (mean of 2.80 with SD=1.07) and password uniqueness in comparison to the other entry methods. Copy&paste exhibits the strongest password uniqueness, however, at the same time the weakest password strength (1.98 on average with a SD=1.33).

2.5.3 Grouping based on creation strategy

We grouped our participants based on their self-reported strategies for creating new passwords (see *Q9*, *Q13*, and *Q15* in Appendix A.1). Based on their answers, we discovered a dichotomous grouping:

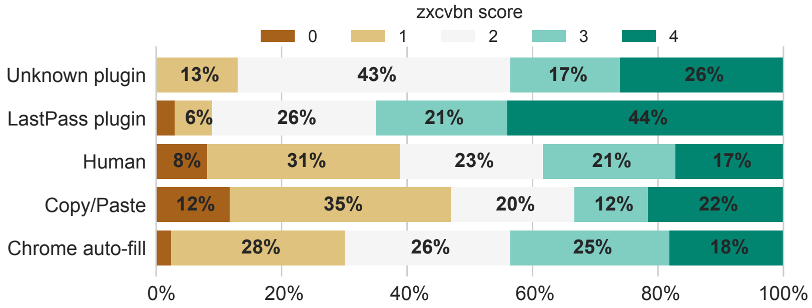


Figure 2.5: Zxcvbn score per entry method for unique passwords.

Group 1: Password managers/generators ("PWM"): First, we identified participants that reported using a password generator, either as integrated part of a password manager program (e.g., *"I use lastpass.com, which automatically creates and saves very strong passwords."*) or as an extra service (*"I use a service to generate/create passwords that I put the parameters in that I would like"*). Many also implied the usage of a manager for password storage (e.g., *"I use a password creation and storage-related browser extension that also is related to an installed password manager application on my personal computer."*), however, some participants explicitly noted a separate storage solution (*"I use an app that creates random character strings to pick new passwords for me. I then memorize it so I don't have to keep it written down"* or *"I will use a random password generator. [...] I will save the new password in a secure location such as a password protected flash drive."*). In total, 45 (or 26.47%) out of 170 participants fell into this category.

Group 2: Human-generated ("Human"): We discovered that all 121 remaining participants described a strategy that abstains from using technical means. Almost all of the participants in this group reported that they *"try to come up with a (random) combination of numbers, letters, and characters."* For instance, one participant symptomatically reported: *"I think of a word I want to use and will remember like. mouse. I then decide to capitalize a letter in it like mOuse. I then add a special character to the word like mOuse@. I then decided a few numbers to add like mOuse@84."* Only a very small subgroup of seven participants reported using analog tools to create passwords, such as dice or books (*"I have a book on my desk I pick a random page number and I use the first letter of the first ten words and put the page number at the end and a period after."*), or using passphrases.

Many of the participants in this group also hinted in their answers to their password storage strategies. For instance, various participants emphasized ease of remembering as a criteria for new passwords (e.g., *"something easy to remember, replace some letters with numbers."*), others use analog or digital storage (e.g., *"I try to remember something easy or I right[sic] it down on my computer and copy&paste it when needed."*). Many participants also admitted re-using passwords as their strategy (e.g., *"I use the same password I always use because it has served me well all these years"* and *"I have several go to words i use and add numbers and symbols that i can remember"*).

	Human	PWM
Number of participants		
	121	49
Gender		
Female	59 (48.76%)	14 (28.57%)
Male	62 (51.24%)	35 (71.43%)
Age group		
18–30	48 (39.67%)	16 (32.65%)
31–40	39 (32.23%)	24 (48.98%)
41–50	27 (22.31%)	5 (10.20%)
51–60	5 (4.13%)	3 (6.12%)
61–70	2 (1.65%)	0
≥71	0 0	1 (2.04%)
Computer science background		
Yes	10 (8.26%)	17 (34.69%)
No	111 (91.74%)	32 (65.13%)
Education level		
Less than high school	0	1 (2.04%)
High school graduate	22 (18.18%)	4 (8.16%)
Some college, no degree	28 (23.14%)	6 (12.24%)
Associate's degree	27 (22.31%)	7 (14.29%)
Bachelor degree	35 (28.93%)	27 (55.10%)
Ph.D	0	1 (2.04%)
Graduate/prof. degree	9 (7.44%)	3 (6.12%)
Ethnicity		
White/Caucasian	91 (75.21%)	32 (65.31%)
Black/African American	15 (12.40%)	10 (20.41%)
Asian	5 (4.13%)	4 (8.16%)
Hispanic/Latino	10 (8.26%)	2 (4.08%)
Multiracial	0	1 (2.04%)
Privacy concern (Westin index)		
Privacy fanatic	45 (37.19%)	21 (42.86%)
Privacy unconcerned	15 (12.40%)	16 (32.65%)
Privacy pragmatist	61 (50.41%)	12 (24.49%)
Attitude about passwords		
Pessimist	1 (0.83%)	2 (4.08%)
Optimist	88 (72.73%)	44 (89.80%)
Conflicted	32 (26.45%)	3 (6.12%)
Prior password leaked experienced		
No	53 (43.80%)	19 (38.78%)
Yes	44 (36.36%)	14 (28.57%)
Not aware of	24 (19.83%)	16 (32.65%)

Table 2.6: Demographics of our two participant categories.

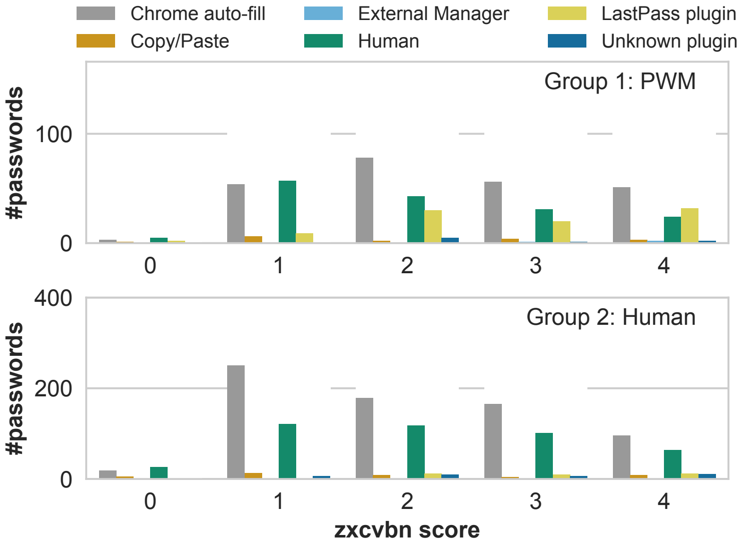


Figure 2.6: Password strength distribution by participant group and broken down by entry method. Hatched bars show the total number of passwords per score.

2.5.3.1 Group demographics

We provide an overview of the groups' demographics in Table 2.6. We again used a Mann-Whitney test to detect any significant differences in the distributions of those two demographic groups. We find that they have statistically significant different distribution for gender ($U = 2,366$, $p = .016$), computer science background ($U = 2,181$, $p < .001$), and attitude towards passwords ($U = 3,440$, $p = .024$). More participants in Group$_{\text{PWM}}$ identified themselves as male in comparison to Group$_{\text{Human}}$. The fractions of participants that have a computer science background and that are optimistic about passwords are higher in the group of password manager users. Gender and computer science background are significantly correlated for our participants (Fisher's exact test: $OR = 3.99$, $p = .005$) as are computer science background and password attitude (chi-square test: $\chi^2 = 9.24$, $p < .01$). One hypothesis for this distribution could be that computer science studies had historically more male students and that their technical background may have induced awareness of the importance of passwords as a security measure and the promised benefits of password managers.

2.5.3.2 Comparison of password strength and reuse

Figures 2.6 and 2.7 provide a comparison of the password strength and reuse between the two groups. The hatched bars indicate the overall number of passwords per zxcvbn score and reuse category. The plain bars break the number of passwords down by entry method. Participants in Group$_{\text{PWM}}$ entered in total 522 passwords and participants in Group$_{\text{Human}}$ entered in total 1245 passwords (both numbers include reused passwords, see Table 2.7). For password strength (see Figure 2.6), neither group contained a noticeable fraction of the weakest passwords (score 0). However, Group$_{\text{Human}}$ shows

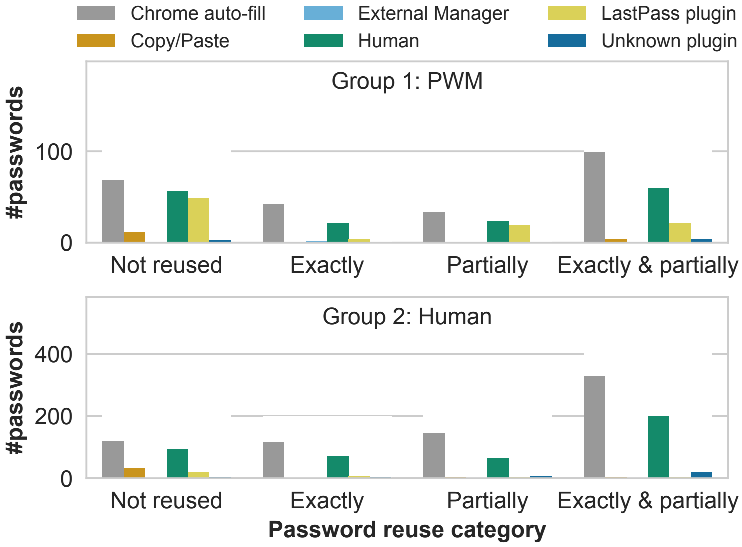


Figure 2.7: Distribution of reuse categories by participant group, broken down by entry method. Hatched bars show the total number of passwords per category.

a clear tendency towards weaker passwords. For instance, there are almost twice as many score 1 passwords ($n = 390$) than score 4 passwords ($n = 191$). In contrast, the most frequent score for Group$_{\text{PWM}}$ is 2 ($n = 158$), but the distribution shows a lower kurtosis (e.g., scores 1, 3, and 4 have the frequencies 126, 113, and 114). When breaking the number of passwords down by their entry method, Chrome auto-fill is the dominating entry method for all zxcvbn scores 1–4 in both groups except for score 1 in Group$_{\text{PWM}}$ where manually entered passwords are most frequent. However, for Group$_{\text{PWM}}$ the fraction of passwords entered with LastPass' plugin ($n = 93$ or 17.82% of the passwords) is considerably larger than for Group$_{\text{Human}}$ ($n = 35$ or 2.81%). In particular, for Group$_{\text{PWM}}$, passwords entered with LastPass have mostly scores higher than 2 ($n = 82$), where score 4 is the most frequent ($n = 32$).

Regarding password reuse (see Figure 2.7), the most frequent category is exactly-and-partially reused ($n = 189$ or 36.21% for Group$_{\text{PWM}}$; $n = 555$ or 44.58% for Group$_{\text{Human}}$). However, Group$_{\text{PWM}}$ shows a bimodal distribution in which not-reused passwords are almost as frequent ($n = 187$) as exactly-and-partially reused ones. Further, Chrome auto-fill is the dominating entry method across all reuse categories in both groups. However, when breaking the passwords down by entry method, more than half ($n = 49$ or 52.69%) of the passwords entered with LastPass in Group$_{\text{PWM}}$ have not been reused in any way. The vast majority of reused passwords can be attributed to manual entry and Chrome auto-fill. In Group$_{\text{PWM}}$, 335 (64.18%) of the passwords have been reused and 979 (78.63%) of the passwords in Group$_{\text{Human}}$. Of the 335 reused passwords in Group$_{\text{PWM}}$, 278 (82.99%) have been entered manually or with Chrome auto-fill. In Group$_{\text{Human}}$, 926 (74.38%) of the reused passwords were entered manually or with auto-fill.

Entry method	Group 1 (PWM)	Group 2 (Human)
All passwords		
Chrome auto-fill	242 (46.36%)	707 (56.79%)
Human	160 (30.65%)	430 (34.54%)
LastPass plugin	93 (17.82%)	35 (2.81%)
Copy&paste	16 (3.07%)	39 (3.13%)
Unknown plugin	8 (1.53%)	33 (2.65%)
External manager	3 (0.57%)	1 (0.08%)
$\sum$	522	1245
Unique passwords		
Chrome auto-fill	144 (42.99%)	396 (55.77%)
Human	101 (30.15%)	230 (32.39%)
LastPass plugin	72 (21.49%)	28 (3.94%)
Copy&paste	14 (4.18%)	37 (5.21%)
Unknown plugin	4 (1.19%)	19 (2.68%)
$\sum$	335	710

Table 2.7: Distribution of entry methods per participant group.

2.5.4 Modeling password strength and reuse

In the next step of our analysis we looked at factors influencing the password strength or password reuse among our participants. Our analyses showed that our participants significantly differ from each other in their average password strength (Kruskal-Wallis one-way analysis of variance, $\chi^2 = 779.19, df = 169, p < .001$) as well as in their average probability of password reuse ($\chi^2 = 692.70, df = 169, p < .001$). The underlying reasons for these differences may be factors that we were able to measure, like the password entry methods of the users, as well as latent characteristics of the users, like their personality or their security awareness. The goal of our further analyses was to show that the effect of the password managers can be shown even beyond these individual differences in password behavior among participants.

One possible way to analyze such a question is a multi-level (aka hierarchical) analysis. This type of regression analysis takes into account the hierarchical structure of our data, where individual password entries are grouped under the corresponding user. Latent, individual differences between users are taken into account in the form of different intercept and/or slope for each user. To get a better understanding of the influencing factors for password strength and reuse, we tested step-wise several regression models. The multi-level models with the studied factors (e.g. entry method) showed a significantly better fit to our data than models that take into account the individual differences between users but do not include the influencing factors we studied. A better fit of the multi-level models was also found in comparison to models that contained the influencing factors but not the individual differences. In the following, we describe our approach to verify the prerequisites for multi-level analysis and our approach to construct the models. Afterwards we report the models for password strength and reuse that fit best to our data.

2.5.4.1 Correlation analysis

Before constructing the models, we started out with a correlation analysis of the available factors (e.g., password composition, participant group, self-reported website value, etc.). As multi-level models are highly vulnerable to multi-collinearity, detecting and potentially removing strongly correlated variables is essential to prevent inaccurate model estimations, which could lead to false positive results. In our dataset, we detected a very high, significant correlation between zxcvbn scores and password composition, in particular password length, as well as with the NIST and Shannon entropies. Since we consider zxcvbn a more realistic measurement of crackability, we omitted NIST and Shannon entropies from our model. Investigation of zxcvbn showed that zxcvbn rewards lengthy passwords with better scores and that its pattern and l33t speak detection can penalize passwords with digits and special characters. Since zxcvbn is the more interesting factor for us and since it partially contains the effect of the password composition on the prediction, we excluded password composition parameters from our models. Moreover, we noticed that password reuse was strongly correlated with the presence of a lowercase character in the password. A closer inspection of our dataset showed, that our data contained a number of PINs, which were all unique, and that every non-PIN password contains at least one lowercase character. In this situation, including the presence/absence of lowercase characters would result in our model just distinguishing between PINs and non-PINs when predicting password reuse.

2.5.4.2 Constructing the models

For both password reuse and strength prediction, we started with a base model without any explanatory variables, which we iteratively extended with additional predictors. In three steps we included a) entry methods, self-reported value, and strength; b) the number of individually submitted passwords per participant, the creation and storage strategy of the user; in a final step c) the interaction between creation strategy and detected entry method. This approach not only allows us to evaluate the effects of the individual explanatory variables, but also to investigate the interplay between different storage strategies and the password creation strategy. In each iteration we computed the model fit and used log likelihood model fit comparison to check whether the new, more complex model fit the data significantly better than the previous one (see Appendix A.5). As our final model we picked the one with the best fit that was significantly better in explaining the empirical data than the previous models. This is a well established procedure for model building, e.g., in social sciences and psychological research [94, 74, 13, 31], and allows the creation of models that have the best trade-off of complexity, stability, and fitness.

2.5.4.3 Zxcvbn score

For the zxcvbn score an ordinal model with all predictors and also the mentioned interaction described our data best. The model is presented in Table 2.8.

The interactions between the self-reported creation strategy (*q9:generator*; see *Q9* in Appendix A.1) and the detected entry methods Chrome auto-fill, copy&paste, and

	Estimate	Std. Error	z value	Pr(>|z|)
em:chrome	0.07	0.12	0.59	0.56
em:copy/paste	-0.13	0.35	-0.89	0.37
em:lastpass	0.24	0.35	0.69	0.49
em:unknownplugin	1.02	0.34	2.97	<0.01
in-situ:value	0.02	0.05	0.48	0.63
in-situ:strength	0.89	0.07	12.68	<0.001
user:entries	0.02	0.02	0.69	0.49
q9:generator	-0.45	0.67	-0.68	0.50
q14:memorize	-0.24	0.30	-0.79	0.43
q14:analog	0.05	0.29	0.16	0.88
q14:digital	0.09	0.31	0.29	0.77
q14:pwm	-0.16	0.28	-0.57	0.57
em:chrome * q9:gen.	2.30	0.60	3.84	<0.001
em:copy/paste * q9:gen.	3.40	1.22	2.79	<0.01
em:lastpass * q9:gen.	1.83	0.82	2.24	<0.05
em:unknownplugin * q9:gen.	0.22	1.34	0.16	0.87

em: Entry method; q9: Creation strategy; q14: Storage strategy; in-situ: Plugin questionnaire

Table 2.8: Logistic multi-level regression model predicting zxcvbn score. Estimates are in relation to manually entered passwords by a human. Statistically significant predictors are shaded. Interactions are marked with *.

LastPass were significant predictors in our model. Those entry methods and also the creation strategy are not significant predictors of password strength on their own. This means that using such a password management tool only leads to significant improvement in the password strength when users also employ some supporting techniques (password generator) for the creation of their passwords. The model might suggest that a general password entry with a plugin (other than LastPass in our dataset) increased the likelihood of a strong password. However, this could be attributed to the high standard error resulting from the minimal data for this entry method.

Moreover, the self-reported password strength was a significant predictor of the measured password strength. This suggests that the users have a very clear view on the strength of the passwords they have entered.

2.5.4.4 Password reuse

For password reuse a logistical model with all predictors but without interactions described our data best. Table 2.9 presents our regression model to predict reuse.

Reuse was significantly influenced by the entry method of the password. In contrast to human entry the odds for reuse were 2.85 time *lower* if the password was entered with LastPass (odds ratio 0.35, predicted probability of reuse with Lastpass = 48.35%) and even 14.29 times *lower* if entered via copy&paste (odds ratio 0.07, predicted probability of reuse with copy&paste = 19.81%). Interestingly, the input via Google Chrome auto-fill even had a negative effect on the uniqueness of the passwords. In contrast to human entry the odds for reuse were 1.65 times *higher* if the password was entered with Chrome auto-fill (odds ratio 1.58, predicted probability of reuse with Chrome auto-fill = 83.72%).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.62	0.45	5.80	<0.001
em:chrome	0.46	0.16	2.81	<0.01
em:copy/paste	-2.68	0.41	-6.54	<0.001
em:lastpass	-1.05	0.37	-2.86	<0.01
em:unknownplugin	0.76	0.51	1.51	0.13
in-situ:value	-0.13	0.06	-2.01	<0.05
in-situ:strength	-0.21	0.08	-2.50	<0.05
user:entries	0.06	0.02	2.67	<0.01
q9:generator	-1.31	0.40	-3.24	<0.01
q14:memorize	0.22	0.25	0.88	0.38
q14:analog	-0.48	0.24	-1.98	<0.05
q14:digital	-0.18	0.26	-0.70	0.48
q14:pwm	-0.07	0.24	-0.30	0.76

em: Entry method; q9: Creation strategy; q14: Storage strategy; in-situ: Plugin questionnaire

Table 2.9: Logistic multi-level regression model predicting reuse. Estimates are in relation to manually entered passwords by a human and refer to the corresponding logit transformed odds ratios. Statistically significant predictors are shaded.

A further significant predictor of password reuse is the user's approach to creating passwords. For users who use technical tools to create their passwords (*q9:generator*), the chances that the passwords are *not* reused are 3.70 times higher (odds ratio 0.27, predicted probability of reuse if technical tools are used = 47.36%). In contrast to the models explaining the zxcvbn-score, our data does not indicate the presence of an interaction effect of the password creation strategy on the relation between entry method and password reuse.

In addition, we found a positive relation between the numbers of passwords entered by users and their reuse. In our model, each additional password of the user *increases* the chance that it will be reused by 6% (odds ratio 1.06). This suggests that with increasing numbers of passwords, it becomes more likely that some of them will be reused, which is in line with prior results [78].

We also found the self-reported website value and password strength a statistically significant predictor for reuse [11]. Passwords entered to a website with a higher value for the user were *less* likely to be reused (odds ratio of 0.87) and also passwords that the users considered stronger were *less* likely to be reused (odds ratio of 0.81).

Lastly, users that reported using an analog password storage (*q14:analog*; see *Q14* in Appendix A.1) were *less* likely to reuse their passwords (odds ratio of 0.62).

2.6 Discussion

2.6.1 Password Managers' Impact

In general, our participants showed very similar password strength and reuse characteristics as in prior studies [154, 196] and our analysis could also reaffirm prior results, such as rampant password reuse.

Our study adds novel insights to the existing literature by considering the exact password entry methods and by painting a more complete picture by considering the users' password creation strategies. We found that almost all participants entered passwords with more than one entry method. Further, we discovered that every entry method showed reused passwords, although the ratio of reused passwords differs significantly between the entry methods. More than 80% of Chrome auto-filled passwords were reused, while only 47% of the passwords entered with LastPass' plugins were reused in some way, and even only 22% of the copied/pasted passwords. Similarly, we noticed that low-strength passwords have been entered with all entry methods, where LastPass had on average the strongest passwords (mean zxcvbn score of 2.80). Interestingly, manually entered passwords and Chrome auto-filled passwords were on a par with the overall password strength but showed above average reuse rates.

For our participants, we discovered a dichotomous distribution of self-reported creation strategies. Participants indicated using a password generator right now or in the recent past, or clearly described mental algorithms and similar methods for human-generated passwords. Taking a differentiated view based on the creation strategies, we find that users of a password generator are closer to a desirable situation with stronger, less reused passwords, although being far from ideal. Only a negligible fraction of participants mentioned analog tools or alternative strategies (like two-factor authentication). Two-factor authentication (2FA), in particular, might be a valuable feature for future, targeted investigations, but for our study, we excluded 2FA since most (major) websites still lack support for 2FA and even for services offering 2FA support the userbase has only little adapted to it [136].

Using regression modeling, we put our data together to a more complete view of password managers' influence. Our models suggest that the interaction between the creation strategy and the entry methods has a significant influence on the password strength. If the passwords are entered with technical support (auto-fill, password manager plugin, or copy&paste), this results in stronger passwords under the condition that technical means were already used when generating the passwords in the first place. Thus, password managers that provide users with password creation features indeed positively influence the overall password strength in the ecosystem. All the more, it is curious that Chrome, as the primary tool to access websites, has the password generation feature disabled by default [93]. Future work could investigate and compare Apple's walled-garden ecosystem, where the Safari browser has this feature enabled by default. Another, maybe surprising, result of our modeling is that the self-reported password strength was a significant predictor for the measured password strength, suggesting that our participants have a clear view on the strength of the entered password. This is in contradiction to prior results of lab studies, like [190], and we think it is worth investigating why users in the wild are so much better at judging their own password strength.

Our models further suggest, that the use of password generators and the website value also significantly reduced the chance of password reuse. More interestingly, however, is that the password storage strategies have different influence independently of an interaction with the creation strategy. Using a password manager plugin or copy&pasting passwords reduced password reuse, while Chrome's auto-fill aggravated reuse. In other words, we observed that users were able to *manually* create more unique passwords when managing their passwords digitally or with a manager, but not with Chrome auto-fill.

The benefit of password managers is also put into better perspective when considering particular strategies in our Group$_{\text{Human}}$. We noticed that users tend to have a "self-centered" view when it comes to password uniqueness (i.e., personal vs. global), but are unaware of the fact that an attacker would not be concerned with *personal* uniqueness of passwords. A large fraction of users reported to *"come up with [a password* ***they****] have never used before"* or to *"try to think of something that [****they****] have never used before."* Those results also align with prior studies [178, 165, 100]. While our participants were able to correctly judge the strength of their entered passwords, their creation strategies indicate an incomplete understanding of uniqueness. In the future, the influence of services like *Have I Been Pwned*[6], which are increasingly integrated into password creation forms and managers, onto the users' understanding of uniqueness and password reuse could be studied.

Another interesting question that comes from our study is why users of password managers (Group$_{\text{PWM}}$) still reuse passwords and employ weak passwords. There could be different reasons, on which we can only speculate at this point. For instance, users might employ a default password for low-value websites, however, we could not find any evidence in our data set for a correlation between website value and strength or reuse for Group$_{\text{PWM}}$. Another explanation could be that those passwords existed prior to starting using a password manager and were never replaced (e.g., LastPass introduced features[7] for automatically updating "legacy passwords" in 2014), or maybe those are passwords that are also required on devices not managed by the user (e.g., computer pool devices at the university). Thus, we think it would interesting to investigate this question more focused.

Further, in light of the high relevance of copy&paste for strong and unique passwords, our results can also underline the "Cobra effect" [97, 98, 138] of disabling paste functionality for password fields on websites to encourage the use of 2FA or password managers. Based on our data, we consider those users who mainly use copy&paste to enter their passwords to be a very interesting subgroup that would be worth further research (e.g., which storage strategies are exactly pursued or motivation to abstain from managers). Unfortunately, there were too few copy&paste users in our current dataset to make any further reliable statements about them separately.

In summary, password managers indeed provide benefits to the users' password strength and uniqueness. Although both benefits can be achieved separately, our data suggest that the integrated workflow of 3rd party password managers for generation and storage provides the highest benefits. More troublesome is that our results suggest that the most widely used manager, Chrome's auto-filling feature, has only a positive effect on password strength when used in conjunction with an additional generator and even shows an aggravating effect on password reuse. The conclusion we draw from this, is that research should investigate how such integrated workflows can be brought to

[6] https://haveibeenpwned.com

[7] https://blog.lastpass.com/2014/12/introducing-auto-password-changing-with.html/

more users, e.g., by better understanding and tackling the reasons why users abstain from using password managers in the first place.

2.6.2 Threats to validity

As with other human-subject and field studies, we cannot eliminate all threats to the validity of our study. We targeted Google Chrome users, which had in general [194] the highest market share, also among our survey participants. Further, we recruited only experienced US workers on Amazon MTurk, which might not be representative for any population or other cultures (external validity), however, our demographics and password statistics show alignment with prior studies. Furthermore, we collected our data *in the wild*, which yields a high ecological validity and avoids common problems of password lab studies [111], but on the downside does not give control over all variables (internal validity). We asked our participants to behave naturally and also tried to encourage this behavior through transparency, availability, and above average payment, however, like closest related work [196, 154] we cannot exclude that some participants behaved unusually.

2.7 Related Work

Textual passwords are for decades [131] the incumbent authentication scheme for online services [89, 91], and will very likely remain in that position for the foreseeable future. They distinguish themselves from alternative schemes through their very intuitive usage, however, as well as through a pathological inability of users to create passwords that withstand guessing attacks [15]. Given the permanence of passwords, users are commonly referred to technical help in form of password management software [85, 183, 168, 178] to create strong, unique passwords.

In this work, we aim to better understand how password managers help users in this task and try to measure the impact password managers actually have on the current status quo. We do this through a comprehensive study that includes both self-reported user strategies and factors for password creation and storage as well as in situ collected password metrics and questionnaire answers. To put our approach into the larger context and to provide necessary background information, we give here an overview of prior research on how users select and (re)use passwords, how password strength can be measured, and on dedicated studies of password manager software.

2.7.1 Password creation

Different works have studied the strategies of users and the factors that influence the selection of new passwords. For instance, users create passwords based on something that has relevance or meaning to them [178], and very often passwords are based on a dictionary word [100, 165].

The effort the user is willing to invest into creating a stronger passwords can depend on different factors. For example, password policies that enforce a certain password composition (i.e., length and character classes) can influence the user [217, 75, 100].

Similarly, many websites use password strength meters to provide real-time feedback on new password's strength and nudge users into creating stronger passwords [61, 189]. However, often those policies and meters have inconsistent metrics across different websites [16, 195, 24], potentially confusing users about what constitutes a strong password [190]. Also the value of the password protected account can influence the user. Prior studies [11, 150, 178, 154] concluded that people try to create strong passwords for accounts that they consider more important, e.g., banking websites. In particular, users employed password managers for specific matters [178], such as just using at a work PC but not at home, or not using them for banking websites. Despite their apparent benefits, it is unclear how users *actually* use password managers and what the exact impact of password managers is on password reuse and strength.

2.7.2 Password strength

Password strength has been studied for several years and different mechanisms have been used to measure a password's strength. Shannon entropy [58] provides a way to estimate the strength based on the passwords composition. It was formerly used by the NIST guidelines [85] to estimate the password strength. However, more recent research [203, 14, 49, 108] argued that *guessability* metrics are a more realistic metric than the commonly used entropy metrics, and recommendations, such as NIST [85], recently picked up the results of this line of research and have been updated accordingly. One of the vital insights from this and other research [96] was that passwords are not chosen randomly but exhibit common patterns and are derived from a limited set of dictionary words.

Measuring a password's guessability has been realized in different ways. Those include Markov models [25, 55], pattern matching plus word mangling rules [205], or neural networks [133]. Since prior password strength meters were based on the password composition and the resulting entropy, those new approaches also found their way into contending password strength meters [205, 133, 188]. However, varying cracking algorithms or techniques can cause varying password strength results based on configuration, methods, or training data [191]. Also in our study we measure the password strength based on guessability, using the openly available *zxcvbn* [205] tool.

2.7.3 Password reuse

Prior work [178] has shown that users have an increasing number of online accounts that require creation of a new password. To cope with the task of remembering a large number of passwords, users resort to reusing passwords across different accounts [37, 99], creating a situation in which one password leak might affect multiple accounts at once. A large-scale data collection through an instrumented browser [75] was first to highlight this problem. Since then, newer studies further illustrated the issue of password reuse. For instance, in a combination of measurement study of real leaked passwords and user survey [37], 43% of the participants reused passwords and often a new password was merely a small modification of an existing one. As with password creation, different factors can influence the password reuse. For example, it was shown that the rate of reused passwords increased with the number of accounts [78], which is troublesome considering that users accumulate an increasing number of accounts. As with password strength, also the value of the website can affect whether a user creates a unique new password or reuses an existing one [11, 154].

Closest to our methodology are two recent studies [196, 154] based on data collected with browser plugins from users. Both studies monitored websites for password entries and recorded the password characteristics, such as length and composition, a participant-specific password hash, the web domain (or domain category), as well as meta-information including installed browser plugins or installed software (e.g., anti-virus software). In case of the newer study [154], also hashes of sub-strings of the password were collected as well as a strength estimate using a neural network based password meter [133] and whether the password was auto-filled or not. Through this data, both studies had an unprecedented insight into user's real password behavior, the factors influencing password reuse, and could show that password reuse, even partial reuse of passwords, is a rampant problem. Further relating to our work, both prior studies also considered the potential influence of password managers, however, could not find any significant effect of password managers on password reuse or strength. However, their studies were not specifically targeted at investigating the impact of password managers, and with our methodology we extend those prior works in two important aspects. First, prior work only considered the presence of password managers and whether auto-fill was used. For our work, we derived a more fine-grained detection of the password entry method, which allows us to distinguish human, plugin-based, auto-fill, or copy&pasted input to password fields and thus better detection of managed passwords. Second, merely the entry method of a password does not reveal its origin (e.g., passwords from a password manager might also be copy&pasted or saved in the browser's auto-fill). To study the impact of password managers, a broader view is essential that includes the users' password creation strategies in addition to their in situ behavior.

2.7.4 Security of password managers

Password manager software has also been the subject of research. Human-subject studies [105, 29] have shown that they might suffer from usability problems and that ordinary users might abstain from using them due to trust issues or not seeing a necessity. Like any other software, password managers might also contain vulnerabilities [127, 220] that can compromise user information. Also the integration of password managers, in particular the password auto-filling, was scrutinized [175, 179] and flaws found that can help an adversary to sniff passwords.

2.8 Conclusion

Passwords are the de-facto authentication scheme on the internet. Since users are very often referred to password managers as a technical solution for creating guessing-resistant, unique passwords, it is important to understand the impact that those managers *actually* have on users' passwords. Studying this impact requires in the first place an approach that is able to detect potential effects of managers. This work's first contribution is an addition to the existing methodology, which for the first time allowed measuring the influence of managers on password strength and reuse *in the wild.* By combining insights into users' password storage and creation strategies within situ collected password metrics, we create a more complete view of passwords. We applied this methodology in a study with 170 workers from Amazon MTurk and were able to show that password managers indeed influence password security. More importantly, we were further able to study factors that affect the password strength and reuse. We found that users that rely on technical support for password creation had both stronger and more unique passwords, even if entered through other channels than a manager. We also found that Chrome's auto-fill option aggravated the password reuse problem. For future work, we see different alleys. For instance, investigating how different, even novel forms of password generators can be integrated with users' strategies. Moreover, one could apply our approach to explore password managers' influence in other ecosystems, such as Apple's walled-garden ecosystem or mobile password managers.

3
FIDO2 Study

Is FIDO2 the Kingslayer of User Authentication? A Comparative Usability Study of FIDO2 Passwordless Authentication

3.1 Motivation and Problem Description

For decades we have tried to replace text-based passwords with more secure alternatives for end-user authentication on the web. But none of the alternatives has achieved this goal until today [15, 90], since none of them could improve security while at the same time offering the same level of deployability and usability as passwords. The newest contender for succeeding text-based passwords is the FIDO2 standard that was jointly developed by the FIDO Alliance—an organization with more than 250 member companies worldwide [70], including Google, Facebook, Microsoft, Amazon, or VISA—and the World Wide Web Consortium (W3C), the main international standards organization for the web. FIDO2 continues the development of the Universal 2nd Factor (U2F) authentication standard and offers websites a standardized way to make use of hardware authentication devices, such as security keys. Like U2F, it supports hardware authentication devices as a second-factor, however, most importantly, it also supports them as a *single-factor* for *passwordless* authentication. Considering the institutions backing FIDO2, this new standard has been presented in the media as a "password-killer" [141, 159, 192, 26]. Also from an academic point of view, using the framework by Bonneau et al. [15] (as we explain in Section 3.3), FIDO2 seems like a promising candidate for succeeding text-based passwords as the incumbent end-user authentication scheme: it provides credentials that cannot be phished, replayed, nor are they subject to server breaches; being an open web authentication standard (WebAuthn), it is supported by virtually all browsers, and native implementations, like on Android and Windows, exist and more are forthcoming; it can provide a consistent user experience; and it supports various authenticator devices, including security keys, like the ones from Yubico or Feitian, but also integrated authenticators commonly available on end-user devices, like Trusted Platform Modules, Android keystore, or Apple TouchID. In fact, in our expert assessment, *none of the existing alternatives to text-based passwords offers as many benefits in Bonneau's et al. framework as FIDO2 with single-factor authentication.*

Thus, while FIDO2 offers strong end-user authentication, high convenience, and has great potential for widespread availability, it is an open question whether end-users accept this paradigm shift from "something they know" to "something they have" (i.e., passwordless authentication). More concretely, we want to find an answer to *"whether end-users accept FIDO2-based authentication as a single factor"* and if not, *"which factors could inhibit an adoption by end-users and which potential paths exist to address the end-user concerns?"*

3.2 Contributions

We conducted the first large-scale comparative user study of FIDO2 passwordless authentication. We recruited 94 participants and randomly distributed them among two groups. In the course of hands-on tasks, one group used a Yubico Security Key as 1FA (passwordless) and the other group, here acting as a control group, used regular text-based passwords for web authentication.Afterwards, we asked participants to reflect on this experience in a survey. The usability and the acceptance of the authentication mechanisms as well as user-specific factors that may effect these variables were measured using standardized methods. In order to get a more complete picture of user perception, we then used free text questions to capture the ideas/benefits/drawbacks/concerns regarding the two authentication methods. As a result, our collected data allowed us to evaluate the usability and acceptance of FIDO2 passwordless authentication and to gather user concerns and feedback about the paradigm shift to FIDO2 passwordless authentication.

Our results show that lay users are very satisfied when directly *replacing* text-based passwords with a security key and are *willing to accept such passwordless authentication over regular text-based passwords.* This is an encouraging result on the road to replace passwords and indicates that FIDO2 has the potential to be the kingslayer of text-based passwords. However, we also identified several potential obstacles that could stop FIDO2 from reaching its goal. Besides known problems of token-based authentication, we identify new issues: First, we find that in case of 1FA, users associate possession of the authenticator with the implicit guarantee that no one else can access the account and, vice versa, the loss of the device with an (impending) illegal account access. This raises the question for a secure and efficient *authenticator revocation* in addition to account recovery—none of which exists as of today. Second, our study identifies new *problems with the physical form factor and features of authenticators.* Our participants questioned the suitability for everyday use and mentioned authentication scenarios for which, in contrast to passwords, they do not see the possibility to use a security key (e.g., public computers without connectivity or delegation of account access to trusted persons). Last but not least, we find that it is often very difficult for users to trust this new technology, mainly because it is such a strong break to previous authentication methods. Our participants had *no mental models* to understand and evaluate the functionality and security of such security keys.

In this light, we find it astonishing that users accept 1FA authentication with security keys so strongly despite these shortcomings. The main reason could be that the disadvantages and weaknesses of text-based passwords have become so obvious and overwhelming for users that they are looking for a technology that can free them from this burden. In summary, we find that there is still a gap between the users' concerns and what the current status-quo of FIDO2 1FA provides. While FIDO2 has the potential to be the kingslayer of passwords, the further development of the standard and of authenticator devices has to more strongly include the perspective of the users and their needs to gain the support of lay end-users. Building upon our results, we try to give concrete recommendations for the supporters of FIDO2, web developers, and further research that hopefully help to foster the proliferation of passwordless authentication on the web.

3.3 Background on FIDO2

FIDO2 is an open authentication standard developed jointly by the Fast Identity Online (FIDO) Alliance and the World Wide Web Consortium (W3C), extending prior work by the FIDO Alliance on the Universal 2nd Factor (U2F) standard, which has also been subject of the academic studies (see Section 3.9). The standard consists

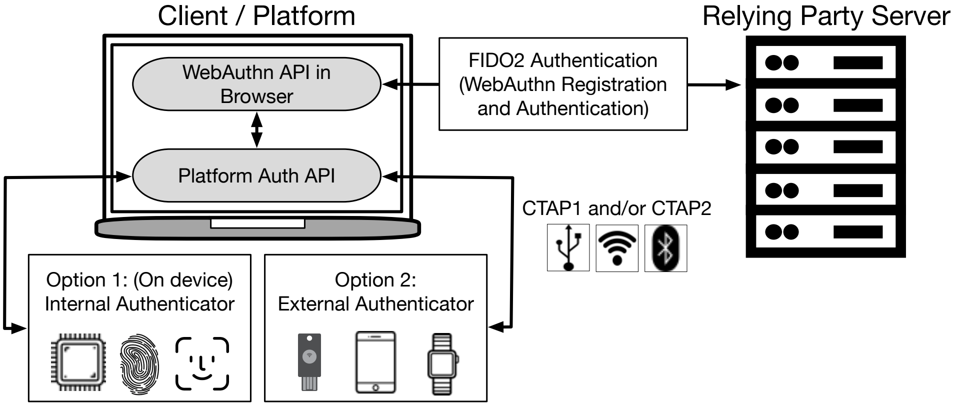

Client / Platform
Relying Party Server
WebAuthn API in Browser
Platform Auth API
FIDO2 Authentication (WebAuthn Registration and Authentication)
CTAP1 and/or CTAP2
Option 1: (On device) Internal Authenticator
Option 2: External Authenticator

Figure 3.1: FIDO2 authentication with WebAuthn and CTAP2.

of two specifications that reflect the two authoring organizations (see Figure 3.1): (1) the WebAuthn protocol [212] for a standardized access by WebAuthn relying parties (e.g., website) to authenticate users via CTAP2 or backwards-compatible via U2F (now considered CTAP1) through a WebAuthn conforming client like the browser; (2) the Client-to-Authenticator-Protocol (CTAP2) [69], an application layer protocol used for communication between a WebAuthn client (like browser) and a conforming cryptographic authenticator device that can either be external and roaming via USB, Bluetooth, or NFC communication (e.g., security key or Android smartphone [68]), or internal (e.g., TPM, Trusted Execution Environment, or TouchID [155]). In contrast to its predecessor U2F, FIDO2 supports two-factor as well as multi-factor and even single-factor (i.e., *passwordless*) user authentication [30]. As a result, FIDO2 supports different levels of user verification, such as a simple test-of-user-presence (e.g., pressing the button on the authenticator) or user authentication to the authenticator via PIN or biometrics. Particularly in single-factor mode, this should ensure user consent to the authentication process.

At the time of writing, various browsers have already integrated stable support for WebAuthn [50], including Chrome, Firefox, Safari, and Edge, and also the number of websites that support WebAuthn is steadily increasing, for instance, Dropbox [81], Microsoft accounts [176, 135], Google accounts, Twitter [210], and others [207] offer FIDO2-based second factors. Also native platform support for FIDO2 is forthcoming, for instance, Microsoft supports it as part of their Windows Hello authentication [132]. Adopters of FIDO2 for non-browser clients or for websites (relying parties) are also supported in their task through an increasing number of FIDO2 libraries and tutorials [18, 135, 218, 5, 157, 158].

In terms of security, FIDO2 is an extension of FIDO U2F and offers the same high security-level based on public key cryptography (see [118] for an overview). At its core, FIDO2 is a challenge-response protocol with mutual authentication using hardware-based authenticators, which offers various advantages over text-based passwords: no shared secrets between user and websites that can be leaked through server breaches, phishing, or key-loggers; unlinkable reuse of the same authenticator for different accounts; or resilience to replay attacks.

Yubico Security Key: The Yubico Security Key is an implementation of a FIDO2 roaming authenticator that offers two-factor, multi-factor, and single-factor (passwordless) authentication. It ships either as a pure USB token or with additional NFC support. It requires neither dedicated hardware (e.g., a reader) nor software, but works with preinstalled drivers on commonly available media (i.e., USB, NFC). To authenticate, users are required to show physical presence during command execution by pressing a capacitive button on the key (i.e., support for test-of-user-presence), indicated by the button flashing. There is no need for any further user input. In our study, we use the USB-only version of the Yubico Security Key as an authenticator in passwordless authentication (see Section 3.6).

Comparison of passwords and FIDO2 1FA with Security Key

We provide context for FIDO2 by applying the framework of Bonneau et al. [15] in an expert assessment to compare the FIDO2 standard to text-based password authentication. Lang et al. [118] also provide a comparison of U2F Security Keys with text-based passwords using this framework and Das et al. [42] concurred with their assessment; however, as we explain in the following, we extend this comparison to FIDO2 and also consider the type of authenticator device as an additional dimension in our assessment.

Bonneau's et al. framework contains 25 subjective factors ("benefits") for measuring the security (11 benefits), deployability (6), and usability (8) of authentication schemes, which also pick up prior recommendations by Stajano [177] for token-based authentication. Table 3.1 summarizes the comparison of benefits that each scheme provides in those categories. As mentioned before, a user could use various types of authenticators, such as USB token, TPM, smartphone, etc. Thus, to apply the framework by Bonneau et al. [15], we had to consider that there are some benefits that only depend on the FIDO2 standard and benefits that are only dependent on the specific design of the authenticator device. This is motivated by the fact that the user primarily has to handle the authenticator, while not being directly concerned with the underlying protocols.

Hence, when we apply the framework by Bonneau et al. [15], we make this explicit distinction between benefits that are derived directly from the FIDO2 protocols and are fixed for all types of authenticators (marked with 🟧 background in Table 3.1), and benefits that are mostly or purely dependent on the authenticator, here, the Yubico Security Key that we used in our study (no background color in Table 3.1). Thus, those benefits might look different if we would use another authenticator, like a smartphone or Apple's TouchID. Here, we only give a summary of our evaluation of FIDO2, a more detailed explanation can be found in Section 3.4.

Summary

FIDO2 with a Yubico Security Key as an authenticator scores almost perfectly in the framework by Bonneau et al. [15], missing *Nothing-to-Carry*, *Easy-Recovery-from-Loss*, *Server-Compatible*, and *Resilience-to-Theft*. In fact, none of the existing alternatives to text-based passwords offers as many benefits in Bonneau's et al. framework as FIDO2 with single factor authentication. While this seemingly makes FIDO2 a very strong

	Usability								Deployability						Security										
Scheme	Memorywise-Effortless	Scalable-for-Users	Nothing-to-Carry	Physically-Effortless	Easy-to-learn	Efficient-to-Use	Infrequent-Errors	Easy-Recovery-from-Loss	Accessible	Negligible-Cost-per-User	Server-Compatible	Browser-Compatible	Mature	Non-Proprietary	Resilient-to-Physical-Observation	Resilient-to-Targeted-Impersonation	Resilient-to-Throttled-Guessing	Resilient-to-Unthrottled-Guessing	Resilient-to-Internal-Observation	Resilient-to-Leaks-from-Other-Verifiers	Resilient-to-Phishing	Resilient-to-Theft	No-trusted-Third-Party	Requiring-Explicit-Consent	Unlinkable
Password	○	○	●	○	●	●	◐	●	●	●	●	●	●	●	○	◐	○	○	○	○	○	●	●	●	●
1FA	●	🟧●	🟧○	●	●	●	●	🟧○	●	◐	🟧○	🟧●	🟧●	🟧●	🟧●	🟧●	🟧●	🟧●	●	🟧●	🟧●	○	🟧●	●	🟧●

● = offers benefit; ◐ = almost offers benefit; ○ = does not offer benefit
🟧 = depends only on FIDO2 standard and is fixed for all authenticators; otherwise, depends purely or mostly on the authenticator device

Table 3.1: Comparison between FIDO2 single-factor authentication using Yubico Security Key and text-based passwords based on the framework by Bonneau et al. (15)

candidate to replace text-based passwords, we are interested in our study in reasons beyond those 25 factors that might affect the acceptance of FIDO2 by users.

3.4 Details on Comparison of Text-based Passwords with FIDO2 1FA using a Yubico Security Key

3.4.1 Benefits dependent only on FIDO2

Usability: FIDO2 is *Scalable-for-Users* (🟧●), as a single authenticator can be used for hundreds of accounts. It never offers the *Nothing-to-Carry* (🟧○) benefit, since it uses *hardware-based* authenticators. FIDO2 does not inherently provide *Easy-Recovery-from-Loss* (🟧○), but the website has to offer recovery or secondary authentication options.

Deployability: FIDO2 is not *Server-Compatible* (🟧○), since the relying parties have to support it separately and cannot piggyback on text-based password authentication. However, with WebAuthn as a W3C standard implemented in all browsers, FIDO2 is *Browser-Compatible* (🟧●), *Mature* (🟧●), and *Non-Proprietary* (🟧●).[1]

Security: FIDO2 is *Resilient-to-Physical-Observation* (🟧●), since it shifts the authentication to a possession-based factor. As a challenge-response protocol based on a public key cryptography, it is *Resilient-to-Targeted-Impersonation* (🟧●), *Resilient-to-Throttled-Guessing* (🟧●), and *Resilient-to-Unthrottled-Guessing* (🟧●). Further, since there is no

[1] Most authenticators in the market are proprietary, although open-source solutions exist and one can build their own authenticator.

shared secret between the user and the relying party, which has to be entered or sent by the user, or stored by the website, and there is mutual authentication between the authenticator and the website, FIDO2 is also *Resilient-to-Leaks-from-Other-Verifiers* (●) and *Resilient-to-Phishing* (●). FIDO2 has *No-Trusted-Third-Party* (●). Lastly, during the registration a new unique key pair is created by the authenticator per account and use of the key restricted to a single origin, which prevents a linking of authenticators and hence linking/tracking of user accounts (based on the used key pair), and makes FIDO2 *Unlinkable* (●).[2]

3.4.2 Benefits dependent primarily on authenticator

Usability: FIDO2 with the Yubico Security Key as single-factor is *Memorywise-Effortless* (●) because the user only needs to press the capacitive button to authenticate, but does not have to remember a secret. This is a good first example where this benefit purely depends on the authenticator. For instance, Windows Hello uses the TPM as authenticator [176]. The TPM is an internal authenticator and does not have any physical entry method, such as a separate keyboard or button, thus the user has to “show presence” and approve authentication by supplying a PIN (in absence of biometric devices), which would make the TPM as a single factor at most *Quasi-Memorywise-Effortless* according to Bonneau et al. [15]. The simple button push makes the Security Key also *Physically-Effortless* (●). The TPM as authenticator with PIN entry would be at most *Quasi-Physically-Effortless.* Inserting the Security Key and pressing a flashing button when prompted by the browser is not more complicated than using text-based passwords and makes the Security Key *Easy-to-learn* (●) and *Efficient-to-use* (●). It is easily conceivable that not all authenticators are necessarily as intuitive as a button press. Assuming proper implementation, FIDO2 should have *Infrequent-Errors* (●) [118], however, this error rate depends on the authenticator. For instance, biometric user authentication to the authenticator might induce more frequent errors [209].

Deployability: The Security Key is *Accessible* (●), since pushing the button does not form a higher hurdle than a password entry. Again, this might change with a different authenticator device (e.g., being able to handle a smartphone). FIDO2 does not impose additional costs per user on the service, however, depending on the authenticator device, the investment by each user varies. With a one-time investment of \$20–\$27 for the Yubico Security Key, we consider this solution as *Quasi-Negligible-Cost-per-User* (◐), since built-in authenticators like TPM or Apple’s TouchID & FaceID would not incur extra costs, while, for example, the Feitian BioPass FIDO2 Security Key costs with \$50 around twice as much.

Security: Since the Security Key has its own physical button, it is *Resilient-to-Internal-Observation* (●). However, authenticators like the TPM, which do not have a trusted path to the user, are still susceptible to internal observation of the PIN/password to use the TPM and have to rely on complex setups like extended authorization policies to

[2] Naturally, accounts could be linked based on information independent of FIDO2, e.g., username, email address, etc.

overcome this limitation. A security key, when used as single factor, is not *Resilient-to-Theft* (○), since possession of the token alone suffices to authenticate. Requiring pushing the button is *Requiring-Explicit-Consent* (●) to authenticate. Internal authenticators without a trusted path might not be able to provide this benefit.

3.5 Research Questions

It was our goal to answer the following research questions:

1. *How do users perceive FIDO2 passwordless authentication in terms of usability?*
2. *Are users accepting FIDO2 passwordless authentication?*
3. *What thoughts and concerns arise in the users' minds when using FIDO2 passwordless authentication?*

To do so, our user study compared passwordless authentication against traditional password-based authentication. In the following section, we develop concrete hypotheses based on prior research findings.

Usability is determined by the users' perception of how well a technology is suited to effectively, efficiently, and satisfactorily achieve their goals. Passwords as the default authentication method on the web can already cover many points concerning usability [15]. However, the two aspects of usability that text-based passwords cannot satisfy —*Memorywise-Effortless* and *Scalable-for-Users*—are particularly important, as average users nowadays have a large number of online accounts [75, P1]. FIDO2 passwordless authentication fulfills these two important requirements and also has the advantage that it is *Physically-Effortless.* Subsequently, we assume that:

> **H1:** *FIDO2 passwordless authentication has a higher usability than traditional password-based authentication.*

The user acceptance for a (technological) system [36] describes factors that, according to the Technology Acceptance Model [47], are direct precursors of the actual usage of a technology in the future. This makes acceptance particularly important if passwordless authentication aims to replace passwords in the long run. The perceived convenience and usefulness of passwordless authentication could lead to a very high acceptance of this technology. On the other hand, users have been accustomed to using passwords for a long time and this extensive previous experience should also lead to a high acceptance of this technology [193]. Since it is not clear which of the two authentication methods should be accepted more, we assume that there are differences:

> **H2:** *FIDO2 passwordless authentication and the traditional password-based method differ in their acceptance.*

Control variables: Prior research has identified several situational and user-specific variables that may also influence users' acceptance. Therefore, we include the following variables in our experimental design to control for their effects: (1) Usability, is one of the most important predictors of technology usage and acceptance [47] and depends heavily on users' preferences and prior experiences [193]. We therefore assume that usability may have an effect on acceptance regardless of the authentication method. (2) Affinity for technology interaction (ATI) describes a person's tendency to enjoy and proactively engage in technology interaction [76, 10]. People with a high ATI should have more fun using a new authentication method and therefore accept it more. (3) Privacy concerns describe users' concerns that can arise if it is not clear what will happen to one's own data [129]. As new authentication technologies, such as FIDO2, are by their very definition related to private information, we controlled for users' individual privacy concerns. (4) A Computer science background—for example, a corresponding degree or course of studies—imparts technical basics and weaknesses of established authentication methods. In order to exclude an effect of this prior education we controlled for such a background.

3.6 Methodology

The core idea of our study was to look at the perception, acceptance, and thoughts of users about FIDO2 passwordless authentication with a security key and compare these to traditional password-based authentication. Thereby, we used a combination of both quantitative and qualitative approaches that are described in more detail in this section.

3.6.1 Study design and procedure

In our user study we used a between-group research design and invited participants to interact with the registration and authentication process of web applications in a controlled (laboratory) environment to gain hands-on experience.

We explicitly decided to let each participant try only one of the two authentication methods in order to avoid that the participants focus mainly on the differences between both schemes. Corresponding contrast effects [170] that could occur in a within-person design might have introduced significant bias into the qualitative analysis of participants' thoughts and concerns. Therefore, we randomly assigned our participants to a study (referred to as $\text{Group}_{\text{1FA}}$) and a control group ($\text{Group}_{\text{Pass}}$), which differed only in the authentication method available to the participants during this hands-on experience.

Members of $\text{Group}_{\text{1FA}}$ could log in with a self-generated user name and a security key. Thereby, we focused on the Yubico Security Key as the authentication device because it is the most popular end-user security key on the market and has already been the subject of studies in the past [164, 42, 118].

Members of $\text{Group}_{\text{Pass}}$ had to create a password in addition to a user name during registration. Thereby, the only password policy in place was a restriction to a minimum length of 8 characters, which corresponds to the lowest possible hurdle according to the NIST password guidelines [85].

At the beginning of the study, participants read the privacy policies and gave their consent. Afterwards, the participants were led to a workplace with a laptop and (in $\text{Group}_{\text{1FA}}$) a Yubico Security Key. The study consisted of a survey with seven stages that guided participants through the entire process in a standardized form (see Figure 3.2):

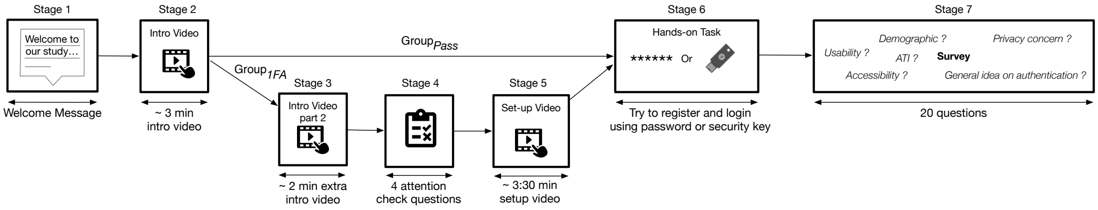


Figure 3.2: Overview of our study procedure.

Stage 1 (welcome message): the study began with a welcome message, including the study instructions.

Stage 2 (topic introduction): Participants watched a video (≈3 min) introducing the topic of the study—"authentication security." From the perspective of Alice (a fictitious character), common problems associated with the registration and use of online services were presented. Alice' story focused on the theft and abuse of account credentials and how to protect against those threats. This video was designed to balance different levels of prior knowledge between our participants.

The next three stages (stages 3, 4 and 5) were only completed by the $\text{Group}_{\text{1FA}}$, while the participants in $\text{Group}_{\text{Pass}}$ were redirected straight to stage 6.

Stage 3 (FIDO2-specific information): Prior work has shown that lack of clarity about the functionality and security benefits of authentication methods leads to lower security ratings, lower acceptance, and reluctance to switch to a new authentication method [202, 201, 204, 42]. FIDO2 is very likely unknown to the users, so we decided to give our participants an introduction to this new technology in order to examine the informed decisions and opinions of users without bias generated by a potential lack of knowledge. Corresponding information was provided to our participants as another video (≈2 min), as it was suggested from users' side in related work [164]. This video dealt with the practical use of a Yubico Security Key for single-factor authentication, its known benefits and drawbacks, and is seamlessly integrated into the introduction video and the story-line of Alice.[3]

Stage 4 (attention check question): Four attention test questions were used to determine if the participants understood the information from the previous stage correctly. None of our participants failed this check.

Stage 5 (setup video:) Afterwards, the participants in $\text{Group}_{\text{1FA}}$ were provided with a setup video (≈3:30 minutes) that explained the setup process for FIDO2 with a Yubico Security Key. The content of the video was a step by step guide through the registration and authentication process using the Yubico Security Key on a demo site that supports FIDO2.

Stage 6 (hands-on task): The participants of both groups received a first-hand experience with their corresponding authentication method. The participants were asked to configure an account on two mockup websites, "Schmoogle" and "Fakebook," which were strongly inspired by the social media service Facebook and the email provider GMail

[3] In practice, most websites do not offer such detailed user guidance for new authentication technologies. In Section 3.7.5, we therefore conduct a "reality check" of our introduction approach to ensure the stability of our results.

to provide a realistic scenario. These two websites were chosen because we assumed that their structure, design, and the way of interaction with them is known to many users. This was especially important as we were not interested in the user interaction with the service as a whole, but especially in the perception of the registration and authentication process. Additionally, there were several reasons why we decided to use mockup websites and not real web services: 1) At the time we designed this study there was no web service that used FIDO2 in passwordless mode; 2) Even though Microsoft is promoting passwordless authentication with FIDO2 for its services [132, 159], a PIN or biometrics is still required to unlock the authenticator, which users may mistake as text-based password or device-local authentication. Moreover, prior works encountered poor user experiences of Windows' support for security keys [164], which we wanted to avoid in our study; 3) We aimed for a controllable and standardized environment, with no risk that our results might be affected by changes in the login process or the user interface of the web service. As a task in $\text{Group}_{\text{Pass}}$ the participants had to register and log-in to the websites using text-based passwords. $\text{Group}_{\text{1FA}}$ had to use the Yubico Security Key to register and authenticate. There was no time limit and participants could try and explore the methods as long as they wanted. For the implementation of our mockup websites for $\text{Group}_{\text{1FA}}$, we used the FIDO2 example projects [157, 158] by Adam Powers. We removed the password fields from registration/login forms. Instead, the registration/login button triggers the WebAuthn API. The introduction and setup videos as well as videos of the workflows for our websites can be found at [79]. Our participants used the Chrome browser for this task.

Stage 7 (survey): After completing the practical task, participants completed the questionnaire with our study variables, which will be described in the next two sections.

3.6.2 Quantitative data collection and analysis

To answer our first two research questions and to test the corresponding hypotheses, we used the following measures. A full overview of all used scales can be found in Appendix B.1.

Usability (SUS). We measured usability ($\alpha = .80$) with the 10-item System Usability Scale (short SUS) from Brooke [21]. Participants stated their level of agreement or disagreement for the 10 items based on their experience with the authentication method. The resulting scores are between 0 and 100 whereby higher scores indicate a higher/better usability.

Acceptance. Acceptance ($\alpha = .90$) was measured with the scale from van der Laan et al. [36]. This scale measures acceptance with 9 semantic differentials. The resulting scores are between 1 and 5 whereby higher scores indicate a higher/greater acceptance.

Affinity for Technology Interaction (ATI). We measured Affinity for Technology Interaction ($\alpha = .92$) using the scale from Franke et al. [76], which measures the construct on a 9-item scale. The resulting scores are between 1 and 6 where higher scores indicate a higher/greater affinity.

Privacy Concern (PC). The participants' privacy concern ($\alpha = .82$) was measured by a 4-item scale taken from Langer et al. [119]. The resulting scores are between 1 and 7 whereby higher scores indicate higher/more privacy concerns.

Demographic Questions. To gain further insight into our study sample, participants answered questions regarding their age, gender, highest educational degree, computer science background, and field of study/work.

3.6.3 Qualitative data collection and analysis

While standardized measuring instruments allow a comparison between our two authentication methods, they are limited in their ability to fully capture individual perception, thoughts, and concerns of users. Therefore, we collected additional qualitative data to answer our third research question.

Our participants answered open-ended text questions about their general impression of the authentication methods, the advantages and disadvantages they see, as well as their willingness to use the method in their personal lives. Our open-ended questions about general impressions were inspired by closest related work [164] and adapted to our specific study setting using best-practices from commercial user experience testing [66] and literature [123] (e.g., recommendations for question form and wording). The open-ended questions about the advantages and disadvantages were added to gain further insights into encouraging and hindering factors in the adaptation of the authentication methods. Additionally, it was of interest to us to find out more about reasons for (un-)willingness to use the authentication methods. The questions were successfully evaluated in a pilot study with five participants, which did not mention any issues. The corresponding questions can be found in Appendix B.1. Subsequently, we used inductive coding (see [123, 80, 208, 134]) to analyze their answers.

In a first step, three researchers independently read all open-ended text answers of our participants and marked all statements that might contain information related to our general research questions. The results were discussed and an initial coding scheme was developed. In the next step, the initial categories were merged by axial coding to category clusters and topics. After this step had been carried out independently by three researchers, they merged their category systems, discussed inconsistencies and created the final code book. Based on this code book, all answers were coded again by two independent researchers. The coders achieved a good mean inter-rater reliability (correspondence between the coders) of Krippendorff's Alpha = .817 [113]. A complete overview of the coding system can be found in Table 3.2.

3.6.4 Ethical concerns

The study design and protocol were reviewed and approved by the ethical review board of our university. We did not collect any personal information, such as username and password. We temporarily stored participants' email address to reimburse them with an Amazon voucher ($12 for ≈45 minutes of participation) and deleted the email addresses after that. All server-side software (i.e., a Limesurvey Community Edition software) was self-hosted on a maintained and hardened university server to which only researchers involved in this study have access.

Topics and aspects
A. Shift from cognitive to physical effort
A.1 Mental effort (password)
A.1.1 Creating passwords
A.1.2 Memorizing passwords
A.1.3 Efficient and easy to use
A.2 Mental effort (1FA)
A.2.1 Reduction of cognitive effort
A.2.2 Efficient and easy to use
A.3 Physical effort (password)
A.3.1 Password entry
A.3.2 No extra hardware
A.4 Physical effort (1FA)
A.4.1 Carrying an extra device
B. Changes in threat model
B.1 Threat model (password)
B.1.1 Cracking and phishing passwords
B.2 Threat model (1FA)
B.2.1 Device theft/loss
B.2.2 Access to account by owner (recovery)
B.2.2 Access to account by other (revocation)
B.2.2 Fallback authentication
C. Restrictions in applicability
C.1 Applicability (password)
C.1.1 Universally applicable
C.2 Applicability (1FA)
C.2.1 Device and connectivity support
C.2.2 Account sharing
D. Breaking with traditions and habitual patterns is hard
D.1 System transparency (password)
D.1.1 Personal secret
D.1.2 Familiar scheme
D.1.3 Positive past experience
D.2 System transparency (1FA)
D.2.1 Mistrust
D.2.2 Lack of knowledge
D.2.3 Perceived security
D.3 Affective perception (password)
D.3.1 Boring / monotonous
D.4 Affective perception (1FA)
D.4.1 Fun / Excitement
D.4.2 Positive feedback about introduction video
E. Security key characteristics
E.1 Robustness and maturity
E.2 Cost

Table 3.2: Code book.

3.7 Results

Our data were collected from mid-December 2018 to end-February 2019 in a laboratory on the campus of our university. Participant recruiting took place via social media groups as well as in lectures and with flyers on our campus.

Variable	$\text{Group}_{\text{Pass}}$	$\text{Group}_{\text{1FA}}$	Statistics	ES
N	48	46		
Gender			$\chi^2(1) = 0.000$	.01
Female	27	26	$p = 1.000$	
Male	20	20		
No answer	1	0		
Age	24.08 (3.63)	25.78 (6.44)	$t(92) = 1.585$ $p = .117$	.33
Education			$\chi^2(5) = 9.462$	.32
< High school	0	2	$p = .052$	
High school	23	12		
Bachelor	12	20		
Master	12	11		
Diploma	0	1		
Ph.D	1	0		
ATI	3.84 (1.12)	4.01 (0.95)	$t(92) = 0.798$ $p = .427$	.16
PC	5.43 (1.31)	5.36 (1.13)	$t(92) = \text{-}0.249$ $p = .804$	.05
CS background			$\chi^2(1) = 4.241$	.23
Yes	18	28	$\boldsymbol{p = .038}$	
No	30	18		
SUS	71.92 (11.09)	81.79 (12.15)	$t(92) = 4.116$ $\boldsymbol{p < .001}$	.85
Acceptance	3.41 (0.70)	4.29 (0.60)	$t\ (92) = 6.522$ $\boldsymbol{p < .001}$	1.35

Note: ES = Effect Size; N = Number of participants; ATI = Affinity for Technology Interaction; PC = Privacy Concerns; CS background = Computer science background; SUS = System Usability Scale. Depending on the variable, the frequencies or the scale mean values including standard deviation are presented in the cells. The statistics column shows the statistical data parameters for a group comparison with two sample t-test respectively with Fisher's exact test for the corresponding variable. p values below the 5% criterion are printed in bold. Effect Sizes are specified in Cohen's d for t-tests and in Cramer's V for Fisher's exact test. N(total) = 94.

Table 3.3: Overview descriptive data.

3.7.1 Sample and participant demographics

Our final sample included N = 94 participants, 56.4% (n = 53) of whom identified themselves as female and the mean age was 24.91. The participants' educational background met the expectations of a university sample. Table 3.3 presents descriptive data for both groups. The second to last column indicates whether there were significant differences between the groups. We found differences for our dependent variables as well as for some control variables which we will discuss in more detail in our statistical analysis. In general, there were no differences in the demographic composition of the groups.

3.7.2 Quantitative results

Usability: Regarding **H1**, an unpaired two-sample t-tests showed significant higher SUS scores in $\text{Group}_{\text{1FA}}$ ($M = 81.74$) than in $\text{Group}_{\text{Pass}}$ ($M = 71.77$); $t(92) = 4.116$, $p < .001$, Cohen's $d = .85$ These results provide support for our hypothesis: **FIDO2**

	Acceptance			
Predictors	*b*	CI	RI	*p*
(Intercept)	3.64	[3.43, 3.84]		**<0.001**
ATI	0.05	[−0.09, 0.19]	1.9%	0.486
PC	−0.01	[−0.10, 0.09]	< 0.1%	0.876
CS (yes)	−0.33	[−0.62, −0.04]	3.7%	**0.025**
SUS	0.02	[0.01, 0.03]	42.8%	**<0.001**
Group (1FA)	0.76	[0.50, 1.02]	51.5%	**<0.001**

Note: Robust regression based on MM estimator [109]. Model 3 can explain 48.8% (R^2adjusted = .488) of the empirical variance (adjusted for number of terms in model); ATI = Affinity for Technology Interaction; PC = Privacy Concerns; CS (yes) = Dummy variable that encodes the effect of a computer science background (No background is the default); SUS = System Usability Scale; Group (1FA) = Dummy variable that encodes the differences for the groups ($\text{Group}_{\text{Pass}}$ is the default). *p*-values below the 5% criterion are printed in bold. N(total) = 94.

Table 3.4: Regression model predicting users acceptance.

passwordless authentication is perceived as more usable than traditional password-based authentication. However, when comparing the SUS scores in our study with other systems and the descriptions provided by Bangor et al. [12] and Sauro et al. [169], both authentication methods are evaluated positively (as "Good", receiving a B grade).

Acceptance: With respect to **H2** an unpaired two-sample *t*-tests showed significant higher acceptance scores in $\text{Group}_{\text{1FA}}$ ($M = 4.29$) than in $\text{Group}_{\text{Pass}}$ ($M = 3.41$); $t(92) = 6.522$, $p < .001$, Cohen's $d = 1.35$. In other words: **Passwordless authentication with the Yubico Security Key was more accepted by our participants than traditional password-based authentication.**

In a next step we assessed the acceptance of the authentication methods with a regression analysis to include the potential effects of the control variables. Stepwise we built regression models including: 1) the control variables, 2) SUS, 3) authentication type ($\text{Group}_{\text{Pass}}$ is the base line against which $\text{Group}_{\text{1FA}}$ is compared) and 4) all possible interactions between all those variables. We used robust regression techniques [110] to calculate the standard error for all estimates in our models, as the Breusch-Pagan test [20] indicated a violation of homoscedasticity ($\chi^2 = 11.949$, df = 5, p-value < .05) for our models. Before the analysis, all metric predictor variables were grand-mean centered to facilitate the interpretation later on.

Considering the complexity of the models, Model 3 containing the first three types of predictor variables mentioned above could explain our empirical data best (see Appendix B.2 for model comparison). This model explained 48.8% ($R^2 = .488$) of the total variance in users' acceptance scores. Table 3.4 gives an overview of the predictors in this model. Our results showed that the SUS score ($b = .02$, $p < .001$), an individual's computer science background ($b = -.33$, $p = .025$), and the predictor representing the difference between the two groups had a significant effect on the acceptance of the authentication methods by the users ($b = .76$, $p < .001$).

Neither ATI nor Privacy Concerns showed a significant effect on the acceptance of the authentication method. A post-hoc relative importance analysis showed that the predictor representing the group differences accounted for the majority (51.5%) of the explained variance while SUS accounted for another 42.8%. Although computer science background is a significant predictor of acceptance, its contribution to the explained variance is very limited (3.7%). The remaining 2.0% can be statistically attributed to the non-significant factors ATI and Privacy Concerns. Overall, these results suggest that: (a) The more usable users perceive an authentication method, the more they will accept that specific authentication method; (b) Even when the control variables are taken into account, the FIDO2 1FA authentication method is widely more accepted than the traditional password-based method; (c) Moreover, we found that people with a computer science background showed in general lower acceptance scores than people without such a background, independently of the authentication method. This is in contrast to recent results about usability of biometrics [209], where experts more readily adopted new technology than non-experts. However, our post-hoc relative weight analysis showed that this effect is minimal and negligible compared to other significant predictors.

3.7.3 Qualitative Results

Qualitative analysis of the free text responses revealed five major concepts for perception, acceptance, and possible use.

Shift from cognitive to physical effort: The vast majority (74; 79% both groups) of our study participants mentioned in one way or another the effort associated with the usage of the specific authentication methods, but in both groups different forms of effort were mentioned. For traditional passwords, primarily the cognitive efforts associated with the use were described. Participants found the creation of secure and unique passwords (5; 10% $\text{Group}_{\text{Pass}}$) but also their memorization (16; 33% $\text{Group}_{\text{Pass}}$) a difficult and demanding task. According to them, the ever-increasing number of accounts that users have to manage are a very burdening factor, as users frequently forget their passwords, resulting in losing access to their accounts.

Regarding passwordless authentication, cognitive effort was not an issue for our participants. In fact, the reduction of cognitive effort compared to password-based authentication was seen as a great (if not the greatest) advantage of passwordless technology (44; 96% $\text{Group}_{\text{1FA}}$).

> *"No recalling of the password. For [a] new account, one need not [to] worry to come up with a password and remember it for later use."*(P92, $\text{Group}_{\text{1FA}}$)

In addition to mental efforts of an authentication method, our participants also described physical efforts associated with these methods. Corresponding topics were particularly evident for passwordless authentication. Eighteen (39%) of the participants in $\text{Group}_{\text{1FA}}$ criticized that this method requires carrying a device to be able to authenticate. It was seen as problematic and annoying that it is not possible to use web services if the security key is not present, which restricts spontaneous and ad hoc use.

> *"I think the only problem with this kind of authentication system is that the user[s] have to carry their Yubikey [Yubico Security Key] everywhere with them [...]"* (P62, $\text{Group}_{\text{1FA}}$)

This physical effort was perceived as one of the major disadvantages of passwordless authentication and led to further concerns which we will discuss later. In contrast to passwordless authentication, only a few of our participants saw a physical effort in classical password-based authentication. Solely the fact that typing passwords can be annoying was mentioned (5; 10% Group$_{\text{Pass}}$) as a disadvantage in this area.

Comparing both authentication methods, the switch from password-based to passwordless authentication was associated with a clear shift in the participants' perception from cognitive to physical effort. This reflects the paradigm shift underlying the switch to FIDO2 1FA—away from "something I know", over to "something I have."

Changes in threat model: Participants from both groups thought about factors and problems that could affect the security of their accounts (59; 63% both groups), although the prevailing threat models differed greatly. In Group$_{\text{Pass}}$ participants (25; 51%) were primarily worried that weak passwords, password reuse, or phishing attacks could lead to an attacker gaining access to their accounts and abusing them.

The participants of Group$_{\text{1FA}}$ (28; 61%) were mainly afraid that someone else could gain access to their accounts with a lost or stolen security key. They were particularly worried as they considered their accounts to be completely unprotected as soon as their key fell into the wrong hands (8; 17% Group$_{\text{1FA}}$).

> *"I just have one concern: What if someone steal[s] my Yubikey [Yubico Security Key]? Does that mean he can access all my accounts just inserting it [to] his computer?"* (P66, Group$_{\text{1FA}}$)

For this reason, some of our participants wanted an additional layer of protection, such as biometrics, to protect the security key against unauthorized use.

> *"[...] I would prefer a finger print verification rather than a push of a button because it is unique only for me."* (P91, Group$_{\text{1FA}}$)

Moreover, our participants (11; 24% Group$_{\text{1FA}}$) were worried about a point that was no issue in Group$_{\text{Pass}}$: The loss of control over one's own account and thus one's own data if the security key is lost, stolen, forgotten, or damaged.

> *"If I forget the YubiKey [Yubico Security Key], I can't get into my accounts."* (P63, Group$_{\text{1FA}}$)

> *"If my Yubikey [Yubico Security Key] gets broken (let's say my coffee spilled on it) I won't be able to login to my accounts."* (P54, Group$_{\text{1FA}}$)

Thereby, several participants raised the question how to "revoke" and "recover" account access in such a case. These concerns went so far that they expressed a desire for a backup authentication method.

> *"There should be a way to use your accounts without the yubikey [Yubico Security Key]. Otherwise you would be very dependent on it."* (P50, Group$_{\text{1FA}}$)

Interestingly, one of our participants, who claims to have *“already been on the receiving end of the password theft,”* points out that the biggest advantage of passwordless authentication is the implicit guarantee that no one else can access users’ accounts as long as they are in possession of their own security key. In this way, the disappearance of the security key from one’s own possession immediately warns the user of a potential (impending) unauthorized access to their account—something that passwords simply cannot offer.

If we compare the two types of authentication, we can see that the threat model for passwordless authentication is fundamentally different from the one for passwords. Because a physical object is required for authentication, the concerns of our participants about threats from the online world, such as phishing or password leaks, are radically reduced. On the other hand, such a dependency brought attention to the inherent natural weakness of such physical objects, their susceptibility to loss, theft, and destruction. Especially the fear of losing access to one’s own accounts seems to be of great concern.

Restrictions in applicability: Another major problem that has arisen in relation to passwordless authentication are situational barriers associated with this type of authentication. Participants (14; 30% $\text{Group}_{\text{1FA}}$) complained about technical incompatibilities, which can be traced back to the specific implementation of the security key, especially the applicability for mobile devices, like smartphones or tablets. For our participants, an implementation using USB, as we studied it, seems problematic and perhaps even outdated.

> *“Nowadays an USB dongle seem to be a bit old, new computer doesn’t have this port, also probably most of the authentication on these days are done in mobile devices. . . ”* (P70, $\text{Group}_{\text{1FA}}$)

On the other hand, participants (7; 14%) from $\text{Group}_{\text{Pass}}$ came up with cases of authentication in which passwords seem to be superior to other technologies because of their flexibility. In this context, they mentioned the ability to spontaneously delegate accounts via telephone or the usage of specially protected computers (e.g., public computer in a library) that do not provide access to standard interfaces.

> *“. . . If necessary, you can also help relatives via telephone or Internet by changing something in their account or doing something for them if they are prevented from doing so.”*(P9, $\text{Group}_{\text{Pass}}$)

> *“Public PCs may not provide an accessible USB interface.”* (P84, $\text{Group}_{\text{1FA}}$)

In summary, these findings indicate that passwordless authentication cannot yet cover all user scenarios (at least with the tested USB implementation) and that neglecting specific corner-cases could be very problematic.

Breaking with traditions and habitual patterns: In contrast to the previous points, many statements of the participants also described aspects connected to the mental migration process from passwords to passwordless authentication. As such, this shift means a break with the well-established habits and traditions of users. Over the course of our study, it became very clear that our participants (40; 82% $\text{Group}_{\text{Pass}}$) have a clear mental model of password-based authentication. They know the pros and cons and have a certain understanding of the factors responsible for the security of a password (35; 71% $\text{Group}_{\text{Pass}}$). At least for our participants this positive mental model does not seem to have been challenged by prior negative experiences (e.g., by account theft) and therefore became the mental default for authentication.

> *"[I use passwords] for all accounts, because I have never had any problems with it, which means my accounts have never been hacked."* (P33, $\text{Group}_{\text{Pass}}$)

For passwordless authentication, on the other hand, such mental models must first be established in the users' minds. Although the videos in our study already seem to be a helpful introduction to this new technology from the participants' point of view (5; 11% $\text{Group}_{\text{1FA}}$), obvious misconceptions in the free-text responses (27; 59% $\text{Group}_{\text{1FA}}$) show that their mental models are only rudimentary.

> *"Is it possible to track my exact location once I insert the Yubikey [Yubico Security Key]?"* (P52, $\text{Group}_{\text{1FA}}$)

Such a lack of technical background knowledge and the associated lack of trust can be one of the biggest obstacles to the adoption of any new authentication method. One of our participants summarizes this quite clearly:

> *"Most people might rather use a password because they better understand and know how it works."* (P72, $\text{Group}_{\text{1FA}}$)

However, these hindering factors for adoption were countered in our study by an affective reaction to passwordless authentication that was very positive. Thereby, the majority of participants (27; 59%) in $\text{Group}_{\text{1FA}}$ described the authentication as a fun, pleasant, and exciting new user experience.

> *"It was overall very nice and pleasant. I found it very intuitive to use."* (P62, $\text{Group}_{\text{1FA}}$)

This is countered by a rather negative affective reaction to password-based authentication (3; 6% $\text{Group}_{\text{Pass}}$), which is described as "monotonous," "boring," and in total "annoying."

In summary, it can be said that due to the lack of mental models and knowledge about the security of passwordless authentication, it might be still a bumpy road to embed this authentication method as a real alternative to passwords in users' minds. Nevertheless, the very positive affective reaction of our participants to passwordless authentication gives us hope that users are ready to replace passwords.

Security key characteristics: After all these mainly conceptual aspects of FIDO2 passwordless authentication, we would like to mention two further points regarding the specific authenticator we used. A few of the participants (7; 15% $\text{Group}_{\text{1FA}}$) mentioned

Category	N(Cat)	Arguments	N(Arg)
Yes	16	Easy/Secure/Memorywise-effortless	3
Yes, but	13	Fear of losing access to own account	5
		Fear of account access by others	4
		Mistrust	3
		Lack of universal access	3
		Costly	1
Rather not	11	Fear of losing access to own account	4
		Mistrust	4
		Costly	3
		Lack of universal access	3
		Annoying to carry extra device	1
No	6	Mistrust	3
		Annoying to carry extra device	3
		Fear of losing access to own account	2
		Lack of knowledge	1
		Fear of account access by others	1
		Costly	1
		Lack of universal access	1

Note: N(Cat) = No. of participants who fell into that category; N(Arg) = No. of participants naming that argument; Total No. of participants in Group$_{\text{1FA}}$: 46.

Table 3.5: Willingness to (not) use passwordless auth.

experiences that may raise doubts about the robustness and maturity of the device. For instance, the form factor of the Yubico Security Key led to ambiguous and misleading situations for our participants.

> *"[I] inserted the Yubikey [Yubico Security Key] into the wrong slot, and later when the message still kept showing, realized that hadn't inserted into the correct slot."* (P92, Group$_{\text{1FA}}$)

> *"Once the Yubikey [Yubico Security Key] didn't react and I didn't know if I had to press it or it's enough to just hold my finger on it."* (P60, Group$_{\text{1FA}}$)

In addition, several participants (10; 22% Group$_{\text{1FA}}$) considered the price of the Yubico Security Key to be very expensive.

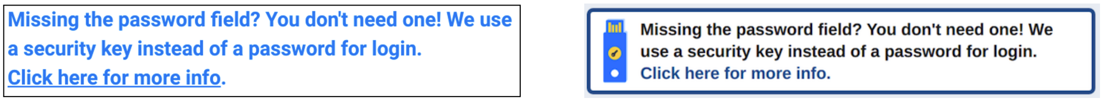
> *"... I don't want to spend money on the key [Yubico Security Key]..."* (P57, Group$_{\text{1FA}}$)

While these findings apply in particular to the security key, we will further address implications and recommendations for the design of authenticator devices in the following discussion.

3.7.4 Willingness to (not) use passwordless authentication

In the end we asked our participants if they now would be willing to use passwordless authentication in their private lives. We identified four different categories in our participants' responses, which we coded as "Yes", "Yes, but", "Rather not" and "No."

Table 3.5 summarizes our participants' answers. We also coded their arguments about why they would (not) use it and we list the most mentioned arguments in the table.



Missing the password field? You don't need one! We use a security key instead of a password for login.
Click here for more info.

Missing the password field? You don't need one! We use a security key instead of a password for login.
Click here for more info.

(a) Hint on Schmoogle

(b) Hint on Fakebook

Figure 3.3: Hint about passwordless login on registration pages with link to modal dialog for information.

Of all 46 participants in Group$_{\text{1FA}}$, 16 (35%) mentioned that they would be willing to use the scheme without any further conditions and explicitly highlighted the ease and convenience of the method over passwords. Most of them also mentioned they would use the method on almost all kinds of websites. This indicates that they found the scheme secure enough to apply even on their most important websites.

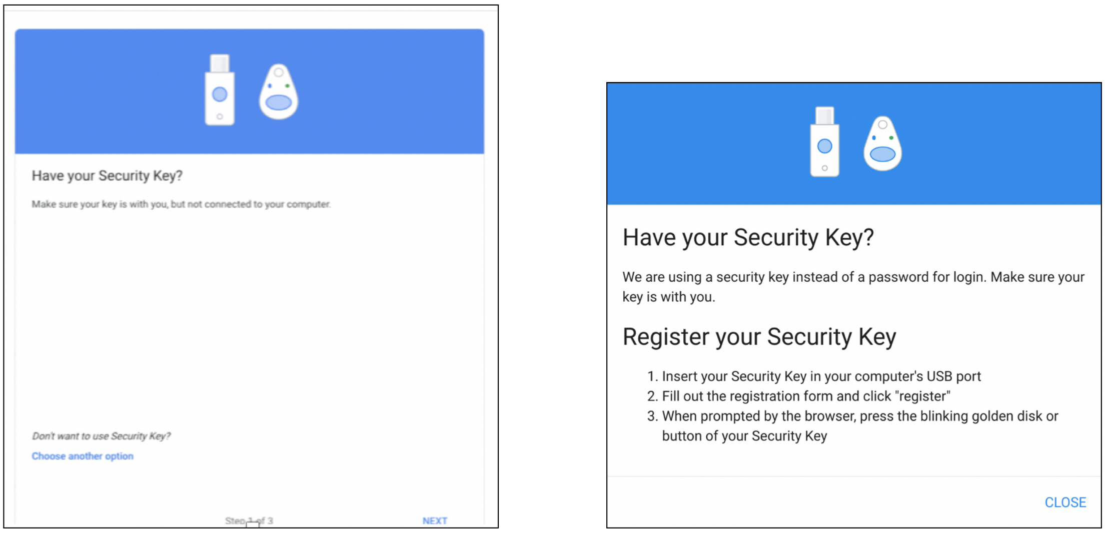
The remaining participants (30; 65% Group$_{\text{1FA}}$) had different kinds of concerns. Participants in the "Yes, but" subgroup gave concrete conditions that have to be met for them to be fully willing to use passwordless authentication, while participants in the "Rather not" and "No" subgroups gave explicit reasons why they are not willing to use passwordless authentication. All three concerned groups mentioned the almost exact same set of arguments, only with slightly different rankings. In general, the "Fear of losing access to the own account" or "Fear of access to their account by others" and "Mistrust" were mentioned most frequently (16; 53% and 10; 33% respectively), followed by "Lack of universal access" and "Costly." Only participants in the "No" subgroup argued more frequently with the "Annoyance to carry an extra device."

Overall, the results in Table 3.5 suggest that there is a high potential willingness to use passwordless authentication over text-based passwords, if certain obstacles were addressed. On the other hand, there are also reasons that seem to discourage users from switching to passwordless authentication. In the following Section 3.8, we make suggestions how most of these problems could be addressed in a way so that passwordless authentication may appeal to the majority of the users.

3.7.5 Stability of Findings

In practice, the process of introducing users to new authentication methods is usually not as detailed as in our study. On the one hand, most websites only offer minimal information in the form of an abstract text and rarely a step-by-step guide. On the other hand, not all users are willing to spend several minutes watching an introduction video. To ensure the validity of our findings also for such conditions, after our main study we tested another group of participants (1FA control group, or short: Group$_{\text{1FAcon}}$) to whom we explicitly provided no detailed introduction about FIDO2 and the security key.

Group$_{\text{1FAcon}}$ ($n = 47$) went through the same test procedure as Group$_{\text{1FA}}$ from our main study except for the following two changes: 1) we omitted the introduction video and any communication of benefits or risks (Stages 2–5 in main study); and 2) we added minimal guidance on how to use the security key in a modal dialog on the websites' registration pages. This dialog was optional for registration/login and only appeared if



Have your Security Key?

Make sure your key is with you, but not connected to your computer.

Don't want to use Security Key?

Have your Security Key?

We are using a security key instead of a password for login. Make sure your key is with you.

Register your Security Key

1. Insert your Security Key in your computer's USB port
2. Fill out the registration form and click "register"
3. When prompted by the browser, press the blinking golden disk or button of your Security Key

(a) Google

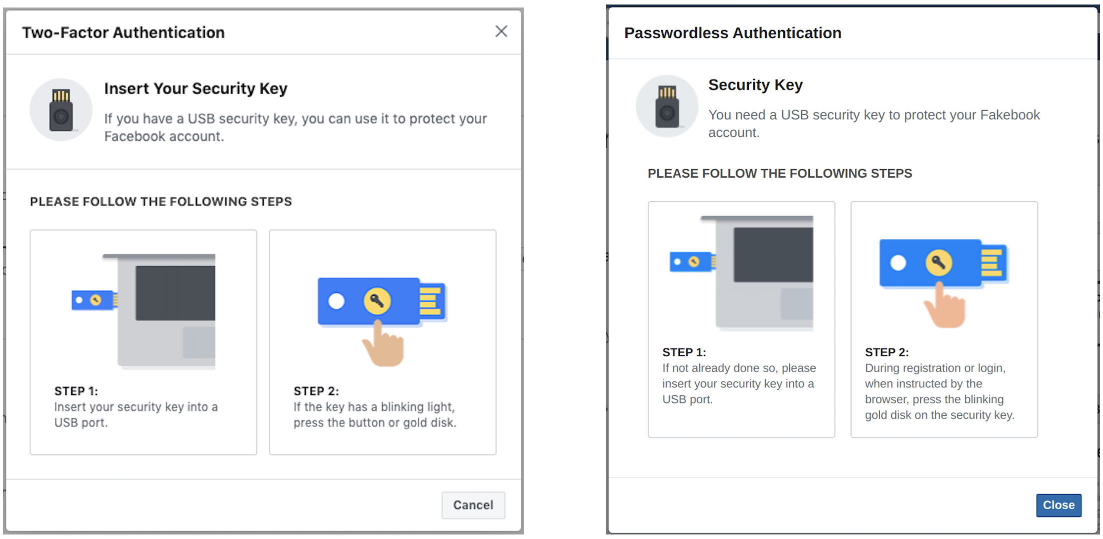
(b) Schmoogle

Figure 3.4: Google vs. Schmoogle.

participants explicitly *"clicked for more info"* on the registration page (see Figure 3.3). The design of this dialog was copied from the 2FA instructions for activating a security key on the actual Facebook and Google sites (see Figures 3.4 and 3.5for a comparison).

Quantitative results: Appendix B.3 provides all analyses presented for the main study supplemented by the data of Group$_{\text{1FAcon}}$. In general, Group$_{\text{1FAcon}}$ did not substantially differ from the other groups in terms of demographic composition. In line with the results from the main study, we found significant higher SUS and acceptance scores in Group$_{\text{1FA}}$ and Group$_{\text{1FAcon}}$ than in Group$_{\text{Pass}}$ ($M = 71.77$) , but no differences between the two FIDO2 groups (Group$_{\text{1FA}}$ and Group$_{\text{1FAcon}}$). A regression analysis, following the approach from our main study, showed very similar results. In total 42.6% of the empirical variance in acceptance could be explained by the predictors in the model. Significant effects on the acceptance were found only for SUS ($b = .03$, $p < .001$) and the predictors that represent the differences between Group$_{\text{Pass}}$ ($b = .70$, $p < .001$) and Group$_{\text{1FAcon}}$ ($b = .66$, $p < .001$). A more detailed analysis showed no significant difference between the two FIDO2 groups ($b = .04$, $p = .720$). In contrast to the main study, a post-hoc relative importance assigned SUS a slightly higher relative importance (56.2%) than the predictors that represent the differences between the groups (41.9%). Thereby, the calculation of the relative importance of predictors is also subject to effects of sampling measurement error, which may explain deviations in this range [104]. In summary, the quantitative results of Group$_{\text{1FAcon}}$ suggested that even without a detailed introduction, FIDO2 passwordless authentication was perceived as more usable and was more accepted than traditional password-based authentication.

Qualitative results: Two independent researchers evaluated the free text answers of Group$_{\text{1FAcon}}$ and neither found a topic that was not yet included in the code-book from



Two-Factor Authentication

Insert Your Security Key

If you have a USB security key, you can use it to protect your Facebook account.

PLEASE FOLLOW THE FOLLOWING STEPS

STEP 1:
Insert your security key into a USB port.

STEP 2:
If the key has a blinking light, press the button or gold disk.

(a) Facebook

Passwordless Authentication

Security Key

You need a USB security key to protect your Fakebook account.

PLEASE FOLLOW THE FOLLOWING STEPS

STEP 1:
If not already done so, please insert your security key into a USB port.

STEP 2:
During registration or login, when instructed by the browser, press the blinking gold disk on the security key.

(b) Fakebook

Figure 3.5: Facebook vs. Fakebook.

the main study. Consequently, this coding scheme was used to allow comparison to the results of the main study. In general, there were only very limited differences in the response patterns between Group$_{\text{1FAcon}}$ and Group$_{\text{1FA}}$. For instance, in both groups a similar proportion of participants mentioned the reduction of cognitive effort as a great advantage of passwordless technology (Group$_{\text{1FAcon}}$ 94%, Group$_{\text{1FA}}$ 96%), but also specific restrictions in applicability of passwordless authentication were mentioned by participants from both groups (Group$_{\text{1FAcon}}$ 13%, Group$_{\text{1FA}}$ 30%). However, specific differences between both groups were found for *B.2 Threat model* and *D.2 System transparency*. In contrast to Group$_{\text{1FA}}$ (17%), a higher proportion of people in Group$_{\text{1FAcon}}$ (49%) were worried as they considered their accounts to be unprotected as soon as their security key fell into the wrong hands (*P47: "I am very afraid that the key will be lost and someone else will get access to all my passwords"*). Also, a larger proportion of the participants in Group$_{\text{1FAcon}}$ (49% vs 20% in Group$_{\text{1FA}}$) showed distrust regarding the security key (*P28: "Privacy, how do they collect our data and how much data do "they" have (Who are "they"?)"*). Additionally, participants in the Group$_{\text{1FAcon}}$ more often (47%) explicitly stated that they lack the knowledge to understand and trust passwordless authentication than in the Group$_{\text{1FA}}$ (17%)(*P43 :"[. . .] I would need more information about how it works, to really judge the key"*). In summary, the qualitative results of Group$_{\text{1FAcon}}$ suggested that even without the detailed introduction of passwordless authentication, the same thoughts and opinions were triggered as in the main study. However, the results also showed that, as expected from previous research, a lack of clarity about the functionality and security benefits of authentication methods can lead to more open questions and concerns among users.

Willingness to (not) use passwordless authentication: We applied the same code book (see Table 3.5 in Section 3.7) for the Group$_{\text{1FAcon}}$ responses about why or why

Category	**N(Cat)**		**Arguments**	**N(Arg)**	
	1FA	1FAcon		1FA	1FAcon
Yes	16	8	Easy/Secure/Memorywise-effortless	3	3
Yes, but	13	25	Fear of losing access to own account	5	1
			Fear of account access by others	4	3
			Mistrust	3	10
			Lack of universal access	3	2
			Costly	1	0
Rather not	11	6	Fear of losing access to own account	4	0
			Mistrust	4	2
			Costly	3	0
			Lack of universal access	3	2
			Annoying to carry extra device	1	2
			Lack of knowledge	0	2
No	6	6	Mistrust	3	4
			Annoying to carry extra device	3	0
			Fear of losing access to own account	2	1
			Lack of knowledge	1	1
			Fear of account access by others	1	1
			Costly	1	0
			Lack of universal access	1	2

Note: N(Cat) = Nr. of participants who fell into the corresponding category; N(Arg) = Nr. of participants who named the corresponding argument; Total Nr. of participants in Group$_{\text{1FA}}$: 46, in Group$_{\text{1FAcon}}$: 47.

Table 3.6: Willingness to (not) use passwordless auth including Group$_{\text{1FAcon}}$.

not they would be willing to use 1FA authentication. Our results show that the "Yes, but" subgroup is the largest in Group$_{\text{1FAcon}}$. In contrast to Group$_{\text{1FA}}$ (13; 28%), 25 (53%) out of 47 participants in Group$_{\text{1FAcon}}$ mentioned that they would be willing to use 1FA under some conditions. This is twice as many as in Group$_{\text{1FA}}$. Most of the participants in both Group$_{\text{1FA}}$ and Group$_{\text{1FAcon}}$ mentioned the almost exact same arguments, only with different ranking: in Group$_{\text{1FAcon}}$, almost of a quarter (10; 21%) of the participants named "Mistrust," while only 3 (6%) mentioned this in Group$_{\text{1FA}}$. A detailed comparison of the willingness among the two 1FA groups is presented in Table 3.6.

3.8 Discussion

We discuss the results of our study and make recommendations to try to address users' concerns.

3.8.1 Closer to a Password Killer?

In our expert assessment, FIDO2 with a security key ticks off almost all benefits and our quantitative results also clearly show that end-users consider this solution both usable and convenient, and do accept it more than text-based passwords. So, is FIDO2 the kingslayer for web authentication? While its high acceptance is encouraging for the future, our qualitative results show a gap between the users' demands and concerns and the current status of FIDO2 authentication with hardware tokens. In the following, we discuss the aspects that we find most interesting in more detail and try to outline recommendations on how the users' concerns could be addressed.

Website	**During 2FA Setup**	**User Settings**
Google (regular)	User can chose between different authentication options (Security Key, Google Prompt, Text message/voice call)	Shows hints and warnings about backup authentication; information on various additional factors
Google (A.P.P.)	User needs two security keys (one key as backup)	—
Dropbox	Only SMS/TOTP as factor, optional backup with phone and OTP	Offers additional factors
Github	Only SMS/TOTP as factor, recommended OTP as backup	Offers additional factors
Facebook	Only SMS/TOTP as factor	Offers additional factors
Twitter	SMS or security key	Offers additional factors

During Setup: User requirements for setting up 2FA and information given during registration about additional authentication options. *User Settings*: Information given to user, if user clicks on "Learn more" or searches the account settings after setup.

Table 3.7: User guidance to set up 2FA on popular websites.

3.8.1.1 Recovery at scale

A predominant concern among the participants in Group$_{1FA}$ was the loss of the security key, which they feared would bar them from accessing their accounts. This is in line with prior user study results on 2FA with security keys. Up until today, this issue has not been properly addressed, e.g., the FIDO Alliance recommends as account recovery practice for relying parties [71] to *"strongly encourage account holders to add additional authenticators when the account is created or when the account with no additional authenticator is identified"*, such that users retain account access in case an authenticator is lost or broken. A review of how top websites advise their users to set up fallback and backup authentication mechanisms (see Table 3.7) showed mixed and inconsistent guidance. Most websites only require setup of one second factor but do not enforce a backup factor, with the notable exceptions of Dropbox and Google's Advanced Protection Program.

A new, very likely future challenge for account recovery with FIDO2 1FA (and even with 2FA), in contrast to prior scenarios, will be the scale of the recovery effort. The unlinkable reuse of a single authenticator is considered a strong point of FIDO2 authentication, since the user only needs one device for all accounts. However, if the device is lost, the user has to potentially recover access to *all* accounts for which this authenticator was registered. Unless the user employed the same backup device for all accounts, allowing for an easy switch of the authenticator, the task of account recovery can become burdensome and frustrating, considering that users have an increasing number of accounts [75]. This can potentially impede future adoption of FIDO2 1FA.

Recommendation: *Reusing an authenticator across websites amplifies the risk of losing access to multiple accounts at once. Users have to be supported and guided in strategies for* scalable *account recovery.*

3.8.1.2 Authenticator revocation

A new concern in this setting that a few participants raised is device theft and account access by the thief. Security discussions around FIDO2 and also prior work on 2FA [42] noted that this risk is lower than the risk of being victim of a phishing campaign or server breach, and further, that the thief needs physical access. This is the objective view of a global risk assessment, which is in stark contrast to the users' subjective view we found. We think that those concerns are discarded too prematurely in a discussion of passwordless authentication. Recent results [27] have shown the length to which abusers in intimate partner violence are willing to go or users might have added personally identifiable information to their key [219] that allows linking the key with accounts. It is unclear to which extent passwordless authentication will ease or hamper such targeted attacks (e.g., a physical token might not be as concealable as a memorized password, but passwords can be more easily phished). We think that if the industry does not take these user concerns seriously, FIDO2 will fail as the password replacement.

Recommendation: *The user has to be able to securely revoke access to their account without the need to first recover access themselves in order to have a chance of account lock-down before the illegitimate access. Potential inspiration can be drawn from established solutions, such as key revocation in PKI [34] or GPG, or revisiting key sharing as in Pico [177].*

3.8.1.3 Corner cases

Some participants pointed out that the Yubico Security Key cannot be used on devices without an (accessible) USB port. In fact, in contrast to passwords, which can be entered anywhere—the *FROM-ANYWHERE* benefit by Stajano [177]—token-based authentication will currently always have corner cases in which it is not applicable (e.g., public or embedded computers without accessible USB, Bluetooth, or NFC interface). We argue that it is unlikely that this situation changes in the near future.

Recommendation: *Users should be informed about corner cases in which they cannot make use of passwordless authentication, since layman users presumably cannot predict consequences of the combination of client devices and authenticator.*

3.8.1.4 Form and features of the authenticator

A few participants pointed out problems with the authenticator we used in our study, a Yubico Security Key. Most of those concerns were about the limited connectivity and hence lack of support for other client devices (e.g., mobile phone via NFC or Bluetooth). Other concerns were about the price of the device, its robustness and usability, the lack of additional authentication to the authenticator, or more generally about the fact that users have to carry an extra device.

Recommendation: *Since FIDO2 does not define the form of the authenticator, just its capabilities and protocols, this is a great opportunity to tailor authenticator form and features to user demands, maybe avoiding the need to buy and carry dedicated devices and offer personalized authentication.*

For instance, mobile phones have been recognized as attractive second factors, since most users already own one, carry them with them all the time and notice their loss quickly [22, 53], they are increasingly equipped with biometrics, and they support multiple media (NFC, Bluetooth). However, other forms are imaginable, such as wearables, like fitness tracker wristbands. Yet, an interesting question is to which extent the authenticator device type could undermine the security guarantees of FIDO2, for instance, if users lose their phone regularly [185], do not protect access to their device [59, 87], or depend on the battery life of the phone [202].

3.8.1.5 Establishing mental models

Finally, during our study, we noticed that our participants identify "authentication" automatically with "passwords" and naturally did not have a mental model of how passwordless authentication with a security key works, what its benefits and drawbacks are, or its applicability. The results of our Group$_{\mathrm{1FA}}$ and Group$_{\mathrm{1FAcon}}$ show that our introduction video panned out positive—our Group$_{\mathrm{1FA}}$ mentioned the security benefits, ease of use, and acceptance of FIDO2 passwordless authentication while Group$_{\mathrm{1FAcon}}$ had remaining trust issues and misunderstood benefits. Yet, the result of Group$_{\mathrm{1FA}}$ is not ideal. Some participants expressed mistrust into the hardware token, mostly due to a lack of transparency, and recent security incidents [17] could reinforce such mistrust. Thus, work that increases the trustworthiness of the device [46] is important. Further, our participants raised concerns that we did not cover in our video (e.g., recovery and revocation) or that we did not predict (e.g., corner cases).

Recommendation: *Transition to FIDO2 passwordless authentication requires establishing mental models of users that see authentication more systematically, drawing from existing models about physical keys (e.g., possession of key means no other can access the account; spare keys can & should be used; do not store them with personally identifying information; associate account and the right physical key; etc.).*

3.8.2 Threats to Validity

Our participants were comparatively young, which is a common problem of lab studies in a university setting. On the other hand, the ATI scores, which usually correlate negatively with the age variable, are in our sample comparable to other studies that had a much more diverse age distribution among their participants (e.g., [76]). This suggests that our results should be fully transferable to age-diverse samples.

For our hands-on tasks, we used artificial scenarios, since FIDO2 passwordless authentication is not (reliably) supported by any service, and our setup phase is simplistic (i.e., no wizards or user settings, but in-place substituting passwords for the security key on the registration/landing page of our websites). Prior work has identified the setup phase as problematic [164] and recommended to study this phase separately. However, for FIDO2 the used security key was really just plug'n'play and even Group$_{\mathrm{1FAcon}}$ with minimal, optional instructions was able to intuitively use it. Thus, we argue this allowed us to study the larger context of users switching to 1FA and to derive concrete recommendations for future studies and their design of the user registration processes.

We only used one type of authenticator (the Yubico Security Key, as one of the most popular authenticators in the market) and did not collect any behavioral data (i.e. the time required for the login process). Therefore, some of our results may only apply to this particular setting and neglect such objective aspects of usability. Both these choices resulted from our focus on qualitative research questions, such as for users' perceptions of FIDO2 and subjective obstacles for the usage of this technology. Future work could follow a pure quantitative approach, that uses a between-subject design to test the usability and acceptance of different types of authenticators (e.g. different form factors or pin protection) as well as the effects on the time efficiency of the login process. However, prior works have already shown that in general security keys are more efficient [118, 161] than text-based passwords.

3.9 Related Work

We review prior works on the usability and acceptability of single-factor and two-factor authentication schemes.

3.9.1 Related studies of single-factor authentication

Replacing text-based passwords with alternatives is a very active research area and because of space constraints we refer to the excellent related work sections by Bonneau et al. [15] and Stajano [177] for a more comprehensive overview. We focus in the following on selected works that are either conceptually closer to FIDO2 or found widespread deployment.

Stajano [177] proposed Pico for replacing passwords with a hardware token, which shares many design aspects with U2F and FIDO2. For instance, it is based on a challenge-response protocol based on public key cryptography, offers mutual authentication between Pico and the verifier, and considers the user's privacy (e.g., no tracking). In an evaluation of Pico's usability in the wild [6], users appreciated avoiding passwords. Although this field study only had 11 users, this can be seen as encouraging for the acceptance of FIDO2. Additional user concerns were recovery in case of device loss and blocking Pico remotely.

TLS *client* certificates [162] can be used for online authentication. However, Parsovs [153] pointed out that current implementations have a poor user experience and that client certificates allow services to track users. The implementation of FIDO2 avoids those privacy risks and its implementation in browsers is tailored to providing a simpler, less error-prone, and more consistent user experience.

Very recently, Conners and Zappala [34] proposed a certificate-based authentication where client certificates are managed with an authenticator. Their *Let's Authenticate* solution provides appealing features, such as automatic account registration/login, easier account recovery, and privacy protection, but builds on top of a CA that issues client credentials to users in contrast to the decentralized nature of FIDO2.

3.9.2 Related studies of two-factor authentication

The usability and acceptability of two-factor authentication with different forms of second factors, such as OTP tokens, SMS, push messages, or most recently U2F Security Keys, has been studied in different works. Here, we focus on the most relevant works to our study of FIDO2 authentication with security keys for *passwordless* authentication.

3.9.2.1 General two-factor authentication

Two-factor authentication solutions for web services have been studied, for instance, by Strouble et al. [180], Weir et al. [201, 202], Gunson et al. [86], Krol et al. [114], or De Christofaro et al. [48]. Generally, their results showed that users found specialized hardware for authentication burdensome, that users lose said hardware, and that convenience is more important than perceived usability and security for users' willingness to adopt a new authentication technology.

Fagan and Khan [63] studied the general motivation of users to (not) follow common computer security advice, including the advice to use two-factor authentication. They also conclude that users abstain from two-factor authentication to avoid inconvenience and cost. In our study, we are interested in concerns that would impede adoption of FIDO2 *single*-factor authentication.

3.9.2.2 Acceptability and usability of 2FA with Security Keys

Usability and convenience have been key design factors for U2F security keys, such as the Yubico Security Key. Recent studies [118, 42, 38, 164, 40] have focused on the acceptability and usability of U2F security keys and are closest and most informative to our study.

Lang et al. [118] report about the two year experience by Google for deploying U2F security keys to more than 50,000 of their employees. Their results showed that security keys are easy to deploy and refer to their use as "brainless" in comparison to OTP-based two-factor authentication. However, they did not conduct any user study but rely on user feedback and logs (e.g., authentication attempts or time spent authenticating).

Das et al. [42, 38] conducted a two-phase study and asked their participants to setup a U2F Yubico Security Key as second factor for their GMail accounts. Their results showed that clearer setup instructions led to significant improvements in usability, but did not change the overall acceptability of the solution. A major constraint on the acceptability was the concern about loss of the key, where concern about being locked out of the account was more salient than losing access to an attacker. Many of the participants were also confused about how to recover their account in case their key is lost. Their results highlight that the acceptance of the solution does not depend solely on convenience and usability.

Reynolds et al. [164] describe two usability studies of Yubico YubiKey as second factor: setup and day-to-day usage. In the first study, 31 participants were asked to setup and configure the YubiKey for a Windows 10, Google, and Facebook account. The result of the first study revealed that most participants struggled to setup their accounts with 2FA in general and Yubikey in particular. In a follow-up study, 25 participants were asked to use a Yubikey in their daily lives for a four-week period. In contrast to the first study, participants in the second study reported that the Yubikey is usable in day-to-day usage and gave a high SUS [21] score. However, in both studies, participants had consistent problems with using the YubiKey on Windows 10, which also affected our decision to focus on web authentication with two mockup websites instead of using Windows 10, currently being the only platform supporting FIDO2 *single-factor* authentication. Moreover, FIDO2 has been integrated into browser software and platform support exists, which removes many of the problems the participants in Reynolds' et al. study encountered. Reynolds et al. further recommend to standardize the setup process to improve usability of this crucial step. In our study, we improved the setup part by showing a short video to our participants and explaining step by step how to log into accounts using a Yubico Security Key.

Reese et al. [161] conducted comparative usability studies of the usage and setup of five two-factor authentication methods: SMS, push notifications, TOTP, U2F with Security Key, and printed out codes. The goal was to eliminate confounding factors and provide better comparison of these methods. Their results show that users generally find all five different methods usable and the majority of participants considers the extra effort worth the gain in security. A third of their participants, however, noted that they do not always have their second factor available, causing inconveniences.

Ciolino et al. [32] conducted a comparative lab study of the setup process of three different U2F authenticator devices and SMS OTP as well as a diary study on the continued use of one such authenticator. Their results underline that the setup of security keys is a high inconvenience for users due to lacking instructions and guidance, and that particular user interface design choices of the web services or by the vendor of the authenticator contribute to this problem. Their participants also expressed concerns about the form factor of the authenticator, e.g., easier losing smaller devices, breaking larger devices, or recognizing buttons as such.

Das et al. [40] investigated the user experience of Security Keys with ten older adults (>60 years) and found that non-inclusive design and inadequate risk communication resulted in minimal adoption in their participant pool. In particular, the form factor of the authenticator device (e.g., too small to be handled easily in daily use) and device compatibility were found to be crucial. Our results indicate that the form factor of the authenticator and the applicability of FIDO2 authentication are of general concern.

3.10 Conclusion

The FIDO2 standard has great potential to become the successor to text-based passwords for user authentication on the web. To gain insights on whether also end-users would accept this paradigm shift from the traditional knowledge-based factor to the new possession-based factor, we conducted a large-scale lab study. Our participants shared with us their impressions, thoughts, and concerns about using FIDO2 passwordless authentication with a Yubico Security Key.

Our results show that users consider FIDO2 passwordless authentication as more usable and more acceptable than the traditional password-based authentication, but also that concerns remain that impede many users' willingness to abandon passwords.

Most notably, the fear of losing the authenticator is not only connected with account recovery but also with an imminent illegal access to the account and the need for revocation—-a subjective threat model by users that differs from the objective risk assessment of FIDO2. Further, limited applicability and critique of the authenticator devices themselves have been pointed out. Thus, our results highlight new hurdles on the road to replace passwords with FIDO2 1FA. We think that these concerns are rooted in a gap between the user's *personal* perspective onto this new technology and the global view of the FIDO2 designers that might not sufficiently include the users' views. In the end, fulfilling users' subjective needs is what determines the success of a new authentication technology. What would be the point of trying to kill the king if the people would not follow the new ruler? We made some recommendations for the supporters and adopters of FIDO2 in an effort to address the concerns we could identify.

4
2FA Study

A Systematic Study of the Consistency of Two-Factor Authentication User Journeys on Top-Ranked Websites

4.1 Motivation and Problem Description

Would you buy a car where the gas and brake pedals are interchanged? You would probably be able to learn to drive this car safely after some acclimatization period. Still, it would be an experience that is very inconsistent with what you are used to, and you would most likely not continue using such an unpleasant car. Like this everyday example, a consistent user experience is crucial for websites to fit the mental models that users built and avoid unnecessarily increasing the users' cognitive load and friction by forcing them to learn something new. This important best practice has been captured in *Jakob's Law of Internet User Experience* [143, 145, 112] as one of several heuristics for user experience [214, 215] and usability [144] that guide website design.

Striving for consistent user experience has ruled website design for years, evident in the design of, e.g., online shopping, banking, forums, blogs, or streaming services. The same best practices also apply to user authentication as part of the user experience. When it comes to the incumbent authentication scheme on the web today, text-based passwords, the user experience of passwords is highly consistent across different websites, although recent work [126] discovered inconsistent password policies for blocklists, strength meters, and composition when setting passwords on the top websites. Regardless of this inconsistency, text-based passwords are notorious for their security issues. Among the different solutions proposed to strengthen user authentication on the web, two-factor authentication (2FA) has been shown to have a very tangible positive effect on account security [118, 199, 137]. Nowadays, 2FA is frequently recommended to end users to improve their security hygiene [160]. Fortunately, many websites are starting to offer 2FA options to their users [2, 77]. However, previous work [32, 164] demonstrated that users struggled with 2FA when their 2FA journey did not match their expectations or previous experiences and advocated for more standardized procedures. In a survey with 2FA adopters (see Appendix C.1), we found corroborating evidence that inconsistent implementations of the 2FA user journey caused friction for users that lowered the usability of 2FA and led users to refuse 2FA or abandon websites. Unfortunately, up to today, *we have only very few insights about how consistent the user experience of 2FA is across different websites.*

4.2 Contributions

To provide new insights about how websites offer 2FA to their users and how consistent this user experience is across websites, we systematically study the 2FA user journeys on 85 popular websites in this work. More specifically, we want to determine whether these websites consistently follow the same design patterns and strategies to offer 2FA to their users. Or, in other words, we are interested in the external functional consistency of the 2FA user journeys across popular websites.

To approach our research question systematically, we need concrete factors based on which we can compare the different user journeys. Unfortunately, such a list of factors does not exist for two-factor authentication, and there is no common guideline or best practice on how to implement the 2FA user flow on websites. Furthermore, 2FA is a technology that has only started gaining wider adoption among websites in the last couple of years and was hence in many cases not part of the initial website design. Additionally, the 2FA ecosystem is fragmented into various options for 2FA, such as TOTP, WebAuthn, push notifications, SMS, or custom solutions, each with its own setup process, dependencies (e.g., hardware token or app), and benefits/drawbacks in terms of usability and security [15, 161]. For these reasons, it was not a priori obvious which exact comparison factors could describe potentially diverse user journeys on different websites.

To solve this challenge, we devised a methodology to derive a list of comparison factors from open and axial coding of existing user journeys on the 85 websites in our data set. As a result, we created a list of 22 comparison factors that describe the user journey from *discovery* of an offered (promoted) 2FA support during sign-in/registration, to the *education* of the user about the available second-factor options and their *setup* processes, to *usage* and *deactivation* of the chosen 2FA option(s). Based on these factors, we then compared the 85 websites in our data set to identify common design patterns and differences and to highlight beneficial or detrimental patterns for user experience.

Our results show that there is no overarching design pattern for the user journey that most websites follow. Instead, we found the design space to be clustered into groups of websites with very similar patterns, some of those favored by the top websites and others by less popular sites. The only design aspects that almost all websites agree on about 2FA are that it is an optional feature, how it should be called and described, and where it should be found in the account settings. In contrast, for the crucial steps of setting up and using 2FA, we found that websites implement mixed strategies, such as varying numbers of simultaneously supported 2FA technologies, inconsistent presentation of device remembrance options, or varying degrees of feedback to users.

According to UX guidelines, this lack of consistency increases users' cognitive load and should be avoided. However, consistency alone does not guarantee a good user experience. We found that several of the more consistently used design patterns have been described in prior work as problematic for user experience, including non-encouraging descriptions or missing possibilities to personalize the 2FA. We also discovered that the journeys of top websites, like `icloud.com`, are outliers from the best practices in the academic literature. Therefore, our results create a call for action to reinvestigate what constitutes a good overall 2FA user experience, to study whether there is a "gold standard" for implementing 2FA user journeys, or to explore the motivations of website developers to implement specific design patterns.

4.3 Background

4.3.1 Two-Factor Authentication

With two-factor authentication enabled on a website, a user must successfully provide two authentication factors to verify their identity. Almost always, the first factor is a traditional text-based password. For the second factor, there are different technical realizations of knowledge, possession, and inherence factors. Most common [2, 23] are *one-time codes* delivered via SMS text-message, phone call, or TOTP [128] apps, like Google Authenticator, Duo, or custom apps that the user registered with the website; *push notifications* by sending an alert message to a dedicated app on the user's phone that asks the user to confirm a login attempt; and hardware tokens via the *U2F* or *FIDO2/WebAuthn* [213] standards that rely on public key cryptography and challenge-response protocols.

Each of these comes with its own set of usability and security benefits and drawbacks [161]. Important for our work is that a website with 2FA support can offer one or multiple of those 2FA options, may even allow users to set one of those solutions up multiple times, or may enforce a particular order in which they can be set up or used.

A commonly acknowledged problem with two-factor authentication is account recovery when a user loses access to a factor (e.g., a mobile device with the TOTP app is unavailable). Often the strategy to avoid lockout from a 2FA-protected account is to set up a dedicated recovery option, such as printed-out one-time passwords that can replace another 2FA option, or to configure multiple 2FA options, when supported by the website, e.g., multiple hardware security keys.

4.3.2 User Experience

Unfortunately, providing an exact definition of "user experience" is very difficult, as there is no consensus on the exact definition [148, 122, 8, 103]. However, a common topic among the definitions is that UX encompasses the various aspects of user interaction with a product, such as a website. Cooper et al. [35] note that there exist three overlapping concerns for UX: form, content, and behavior. While form and content (e.g., UI design or phrasing) have an impact on usability, this work focuses on behavior (i.e., functionality) and only touches on some aspects of form and content.

To help designers provide the best possible user experience, various best practices and general guidelines have been developed (e.g., books [35, 115, 214, 200, 173] or online resources, such as *Laws of UX* [215], *Nielsen Norman Group* [146], or *Interaction Design Foundation* [101]). Among the earliest are Shneiderman's eight "Golden Rules" for interface design [173, 174] and Nielsen's "10 Usability Heuristics for User Interface Design" [144, 142]. Shneiderman's rules state, for instance, that one should strive for consistency and provide informative feedback to users. Of Nielsen's heuristics, heuristic nr. 4, also known as *Jakob's law of Internet user experience* [145], is the most important for this work and provides the motivation to study the consistency of 2FA user journeys across websites. This heuristic states that *"users spend most of their time on other sites"* and that *"users prefer a site to work the same way as all the other sites they already know."* As a consequence, one should *"design for patterns for which users are accustomed."* Having such conventions and consistency helps users build upon existing mental models and avoid cognitive friction by forcing them to learn something new [214]. If an unconventional website mismatches the user's mental model, the website will be difficult to learn, difficult to use, or even rejected [200]. One way to drive *external* consistency is to make ample use of guidelines. For instance, for apps there are Google's Material Design Guidelines [84] and Apple's Human Interaction Guidelines [9]. We are not aware of any general guidelines for implementers and designers of two-factor authentication on websites, although there exist case-specific guidelines (for example, FIDO2 [72]) or small collections of best practices (e.g., [51, 198]).

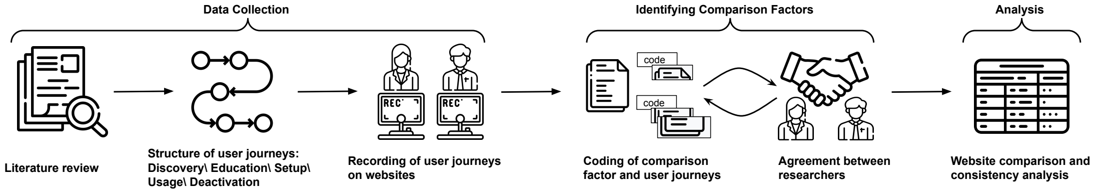

Data Collection
Identifying Comparison Factors
Analysis
Literature review
Structure of user journeys: Discovery\ Education\ Setup\ Usage\ Deactivation
REC
Recording of user journeys on websites
code
Coding of comparison factor and user journeys
Agreement between researchers
Website comparison and consistency analysis

Figure 4.1: Overview of our methodology.

Although in this work we focus on external, *functional consistency*, some of the comparison factors for 2FA user journeys that we identified (see Section 4.6) also touch on other UX guidelines and best practices. Tesler's law [214] states that for any system there is a certain amount of complexity that cannot be reduced, and it is recommended that the product design ensures that as much as possible of the burden on the user is lifted. Krug [115] recommends that if a difficulty for the user cannot be avoided, the design should provide brief and timely guidance, and Cooper et al. [35] recommend contextual help and assistive interfaces without the need to break the user's flow. If it cannot be avoided that the user has to learn something new, users learn best from examples (e.g., pictures, screenshots, or short tutorial videos) [200]. In addition, Hick's law [214] recommends breaking down complex tasks into smaller steps to decrease the cognitive load. Moreover, excise tasks, such as navigational excise, should be reduced, e.g., by reducing the number of places that a user must go and providing clear overviews [35]. Hereby, it is important to consider that users do not read but scan webpages [115] and that this scanning is based on the mental model they built from past experiences, which creates expectations of what they want to see and where [200]. Furthermore, part of Postel's law [214], similar to Shneiderman's third golden rule [173, 174], recommends providing clear feedback to users, and the Peak-End Rule [214] recommends paying attention to the final moments of the user journey because people judge an experience largely based on how they felt at its peak and recall negative experiences more vividly than positive ones. Lastly, personalization can enhance the user experience. Although we did not explicitly investigate websites for their quality of those additional guidelines, some of our comparison factors indicate if 2FA settings are found in common places, if additional information and instructions are provided, if user notifications are present, or if users can set preferences.

4.4 Methodology

To compare the 2FA user journeys of different websites and measure their consistency, we require concrete comparison factors that describe these journeys. Unfortunately, there is no existing list of such comparison factors, of which we are aware, or general guidelines for implementing 2FA on websites from which we could extract such factors. Therefore, a crucial challenge for our study is to create a list of relevant and representative factors. We used inductive research methods (e.g., [124, Chapter 11.4]) to solve this challenge. Figure 4.1 gives an overview of our methodology, whose data collection (Section 4.4.1) and identification of comparison factors (Section 4.4.2) we explain in the following. In a nutshell, we use open and axial coding from grounded theory on the screen-recorded 2FA user journeys of different websites to identify the list of comparison factors and to form an agreement about how each website matches each factor. Using the coding results, we then compare the different websites and study how consistently they implement the 2FA user journey and where they differ (Sections 4.6 and 4.7).

4.4.1 Data Collection

The first part of our methodology is to collect a representative data set of user journeys recorded on different websites that we can analyze. Since we are building our knowledge about user journeys inductively, the screen recordings must have as high as possible coverage of all steps and choices along each journey. To this end, an automated tool, such as a Web crawler, could be used to explore various websites. Unfortunately, the need for a prior knowledge about how websites might implement their user journeys to guide the crawler and the need to use additional authentication devices (e.g., phone or security key) hamper an automated collection. Alternatively, we could use a crowd-sourced data collection, e.g., Amazon Mechanical Turk. Unfortunately, this was not possible in our setting for ethical reasons. We would need to ask our participants to use private accounts (or create fake accounts) on different websites and explore security settings for which they might need to provide a (personal) email address, phone number, or security key, and risk accidentally locking themselves out of an account as a result of a misconfiguration.

Instead, two researchers independently explored and screen-recorded the 2FA user journeys for our study. Their general instruction was to "thoroughly explore all aspects" of these journeys. However, this exploration could be informed beforehand from the literature, which discusses different aspects of 2FA user journeys (see also Section 4.9). For example, recent works (e.g., [82, 120, 41]) and guidelines [72] identify discovery of 2FA options and user education, different works studied 2FA setup and login (e.g., [161, 164]) or mandating 2FA (e.g., [4]), and account recovery is a commonly identified problem. Based on those insights, we structure the exploration of user journeys into five steps: The first step is *Discovery* of 2FA support on the website. We explore the landing pages, FAQ, and account registration for information on 2FA and follow all linked information. To further encourage users, there might also be nudges and messages about securing the account with 2FA, for which we scan the websites' interfaces. To use 2FA, the user must find the corresponding settings in their account settings, which we explore for the locations and options for authentication. In the next step, *Education*, we examine how a website introduces 2FA and if it gives further explanations, such as descriptions of how 2FA works and what it offers. Once the user has decided to use 2FA, they need to *Setup* their second factor(s). We explore the workflow of setting up all supported 2FA options (e.g., TOTP or Security Key). This exploration includes examining the websites' instructions, exploring the different settings choices (e.g., personalization choices), and feedback from the website on successful setup. After setting up two-factor authentication, we examine the *Usage* of 2FA on the website. We re-login and observe how the website prompts us to authenticate and whether it provides any options (e.g., device remembrance), which we explore. Finally, we explore the 2FA

Deactivation procedure in the website settings and how the website communicates those changes.

For data collection, we maintained identical study conditions. All recordings were made on MacBooks running macOS 11 in the same network with the latest version of the Chrome browser when we started our data collection. Data collection was carried out between 06/2021 and 08/2021. This fixed setup should minimize the risk [206] of external factors (e.g., varying geolocation) and possible risk-based authentication to distort the data.

It is important to note that we focus *only* on the workflow for account creation, initial 2FA setup, and 2FA usage. We do not explore the workflows for account recovery or to change personal information relevant to 2FA after 2FA setup, such as a phone number or email address. We consider those follow-up problems to be studied after we have insights into the consistency of the fundamental steps that mint the users' first impressions about 2FA on a particular website.

4.4.2 Identifying Comparison Factors

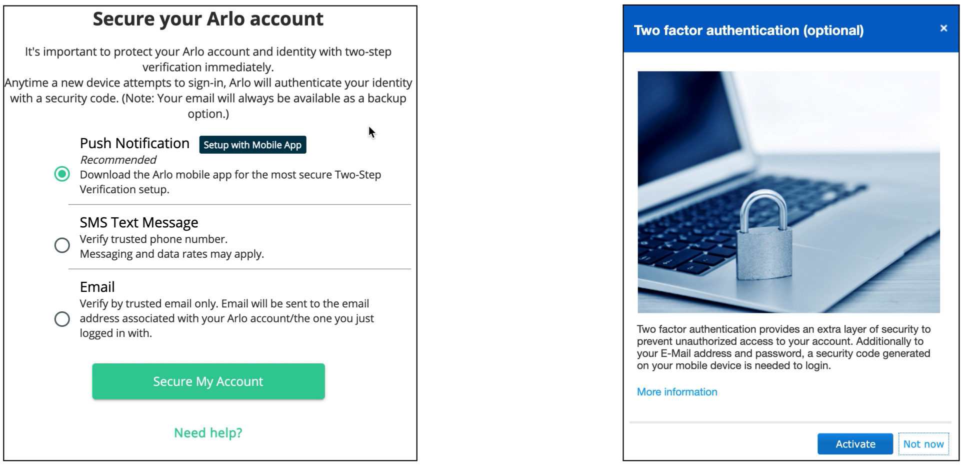
Since there is no predefined set of factors to compare 2FA user journeys, we applied emergent coding [124, Chapter 11.4], in particular open and axial coding from grounded theory, to identify comparison factors from our recorded user journeys. These coding techniques are commonly applied in qualitative data analysis for text content. To still use those established methods, we treated the screen-recorded journeys like semi-structured interviews. Semi-structured interviews follow a set of predetermined questions, but the remaining questions are made up during the interview based on the interviewee's answers. We transferred this idea to our data collection (see Section 4.4.1): The exploration of user journeys follows a set of predetermined questions for discovery over usage to deactivation but allows the researcher to divert to individually explore a website in more detail and discover new or unique aspects of 2FA user journeys. Two researchers separately iterated through the set of recorded user journeys and segmented the observed journeys into meaningful parts to which they assigned concepts (i.e., codes). This is followed by axial coding, where the two researchers combined those concepts via induction and deduction into categories. For example, the codes "2FA advertised on the landing page" and "2FA recommended during account creation" can be combined into "Promotion of 2FA." These combined concepts can be used as comparison factors on all websites. The researchers also noted whether there exists a functional dependency between factors. After agreeing on the list of comparison factors, the researchers discussed how each website matches each comparison factor (e.g., fully, partially, or not at all). Since the matching of comparison factors might reveal that the list of factors is too fine-grained, potentially weighting small differences too heavily, or too coarse-grained, potentially hiding important differences, the researchers repeated the axial coding process until a set of comparison factors and website matching was found to which all involved researchers agreed. The focus of coding was on the *functional* aspects of the websites, and less on the elements of the content or user interface since this study focuses on the consistency between websites and *not* rating the quality of each website's user journey.

4.5 Data Set

To gather a set of websites for our study, we rely on the open source project `2fa.directory` [3, 2] that maintains a list of websites with 2FA support, which almost 1,000 contributors currently curate. The websites are assigned to different categories, such as social, communication, or retail. Since `2fa.directory` distinguishes websites at the level of subdomains, we merged subdomains into their domain when we were aware that they use the same account for authentication. For example, `drive.google.com`, `cloud.google.com`, and `mail.google.com` are in different categories but rely on the same Google account, while `amazon.com` and `aws.amazon.com` have separate accounts. For merged entries, we chose the category we thought end users most likely knew the domain for (e.g., *mail* for `google.com`). Since we rely on a manual investigation of the user journeys of each website, we needed to reduce the set of all websites listed on `2fa.directory` to a feasible number. First, we excluded categories for which we cannot create an account, for example, almost all websites in the *banking* and *government* categories. Second, we used the Tranco [125, 184] data set to rank websites according to their popularity. We selected the top websites from each category, where we selected the number of websites from each category based on the category's weight in the 2fa.directory data set. For example, there were only four *VPN provider* websites in the `2fa.directory` set but 45 *Gaming* websites. This initially resulted in 120 websites. Unfortunately, we had to exclude 35 websites that we could not study for different reasons, such as language barriers, geo-restrictions, or the need for financial expenditures. In the end, we recorded the 2FA user journey on 85 websites with 2FA support from 26 categories.

4.6 Comparison Factors

In this section, we explain the comparison factors that we identified in our analysis of 85 popular websites following the methodology of Section 4.4 and describe informally how we categorize websites according to these factors. We apply the methodology of Bonneau et al. [15] by categorizing every website if it matches (●), partially matches (◐, ◑), or not matches (□) a factor. However, in our categorization, some factors are dependent on other factors, and we denote it explicitly when a conditional factor's prerequisite is not fulfilled (■) and this factor does not apply to a website. Further, in contrast to Bonneau et al., we do *not* use the categorization as a ranking to determine if a website is better than another website, but we use the categorization to identify patterns in how websites realize their 2FA user journey and to study whether websites realize this journey in a consistent way. Although, for some of the factors described below, this categorization overlaps with a scale from known best practices to known poor practices from the literature. We found 22 comparison factors; 8 are conditional and depend on other factors to be applicable.



Secure your Arlo account

It's important to protect your Arlo account and identity with two-step verification immediately.
Anytime a new device attempts to sign-in, Arlo will authenticate your identity with a security code. (Note: Your email will always be available as a backup option.)

● Push Notification
Recommended
Download the Arlo mobile app for the most secure Two-Step Verification setup.

○ SMS Text Message
Verify trusted phone number.
Messaging and data rates may apply.

○ Email
Verify by trusted email only. Email will be sent to the email address associated with your Arlo account/the one you just logged in with.

Two factor authentication (optional)

Two factor authentication provides an extra layer of security to prevent unauthorized access to your account. Additionally to your E-Mail address and password, a security code generated on your mobile device is needed to login.

(a) arlo.com promotes its 2FA as a pop-up immediately after an account creation (*Promotion*: ●) and it is mandatory to setup 2FA for their user account (*Non-optional*: ●). Further, the pop-up provides *Descriptive-notification* (●), *Additional-information* (●), and *Option-specific-information* (●).

(b) teamviewer.com promotes its 2FA as a pop-up immediately after an account creation (*Promotion*: ●) but it is optional to setup 2FA (*Non-optional*: □). Further, the pop-up provides *Descriptive-notification* (●) and *Additional-information* (●).

Figure 4.2: Examples of websites that match the *Promotion* factor and (not) matched the *Non-Optional* factor.

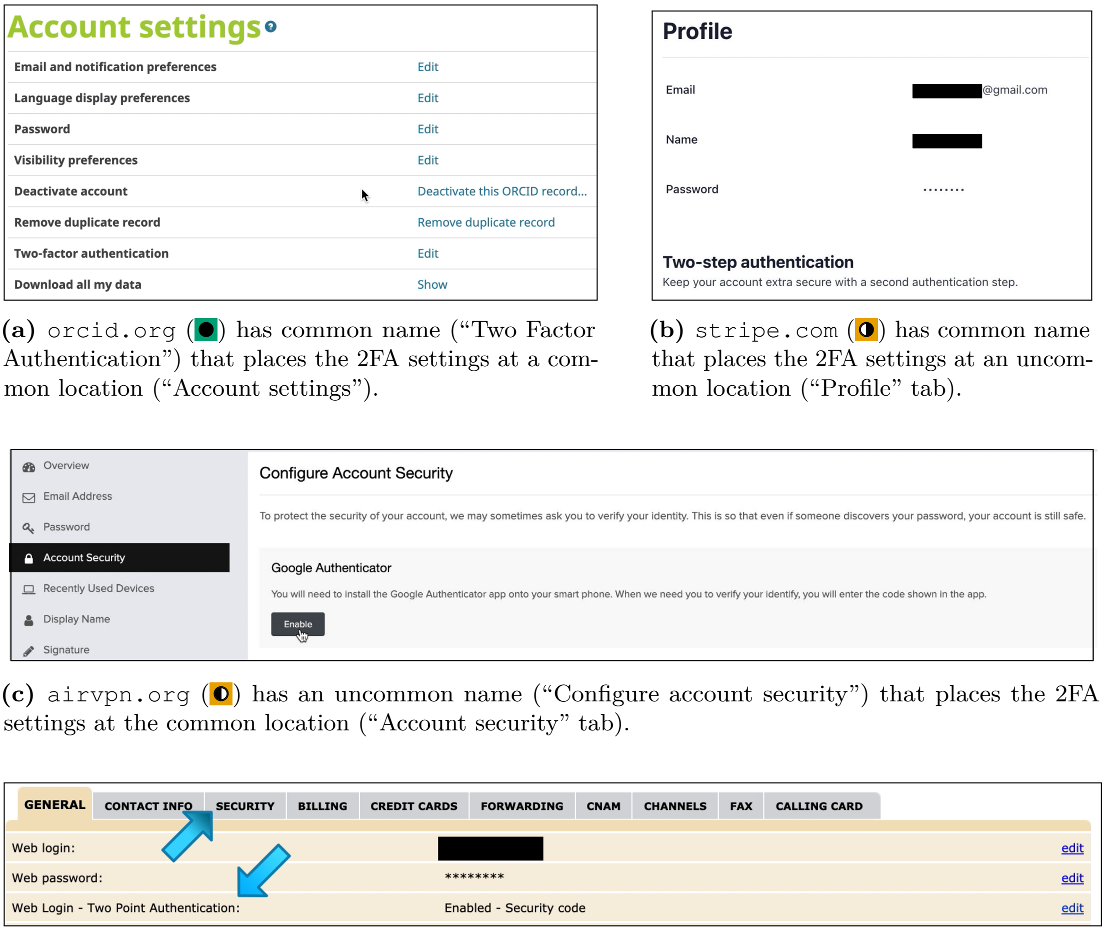
4.6.1 Factors for Discovery

Promotion: The website promotes its 2FA support in a clear and obvious way during account creation or immediately after login (e.g., through a banner, pop-up, or highlighted message) (●). If the website does not clearly promote but only mentions the 2FA support in a way that could be easily missed by the user (for example, only a quick link in the footer of the landing page), we categorize this as *quasi*-promotion (◐). If the service does not promote its 2FA support and the user has to discover it themselves (e.g., browsing the settings pages), we categorize this as not matching (□).

Non-Optional: The website mandates setting up 2FA for user accounts (●). For instance, without setting a 2FA option up, the account registration cannot be completed; or after account registration, core functionality and features of the website are not available to the user until the user sets up 2FA for their account. Otherwise, using 2FA is optional and not mandatory for the website (□).

Figure 4.2 illustrates examples for both *Promotion* and *Non-Optional*.

Common-Naming-and-Location: The website denotes its 2FA settings with a commonly used name, and the 2FA settings are in a commonly used location in the account settings (●). We identify commonly used names and locations in our analysis of our selected websites and summarize the results in Section 4.7.1. If either the name (◐) or the location (◐) is uncommon, we categorize this as *quasi-common-naming-and-location*. If the naming and location are uncommon, we categorize the website as not matching this factor (□) (see Figure 4.3).



Account settings

Email and notification preferences	Edit
Language display preferences	Edit
Password	Edit
Visibility preferences	Edit
Deactivate account	Deactivate this ORCID record...
Remove duplicate record	Remove duplicate record
Two-factor authentication	Edit
Download all my data	Show

(a) orcid.org (●) has common name ("Two Factor Authentication") that places the 2FA settings at a common location ("Account settings").

Profile

Email	[redacted]@gmail.com
Name	[redacted]
Password	········

Two-step authentication

Keep your account extra secure with a second authentication step.

(b) stripe.com (◐) has common name that places the 2FA settings at an uncommon location ("Profile" tab).

Configure Account Security

To protect the security of your account, we may sometimes ask you to verify your identity. This is so that even if someone discovers your password, your account is still safe.

Google Authenticator

You will need to install the Google Authenticator app onto your smart phone. When we need you to verify your identify, you will enter the code shown in the app.

(c) airvpn.org (◐) has an uncommon name ("Configure account security") that places the 2FA settings at the common location ("Account security" tab).

Web login:	[redacted]	edit
Web password:	********	edit
Web Login - Two Point Authentication:	Enabled - Security code	edit

(d) callcentric.com (□) has uncommon name ("Two Point Authentication") for 2FA that places the 2FA settings at an uncommon location ("General" tab) despite the dedicated "Security" tab.

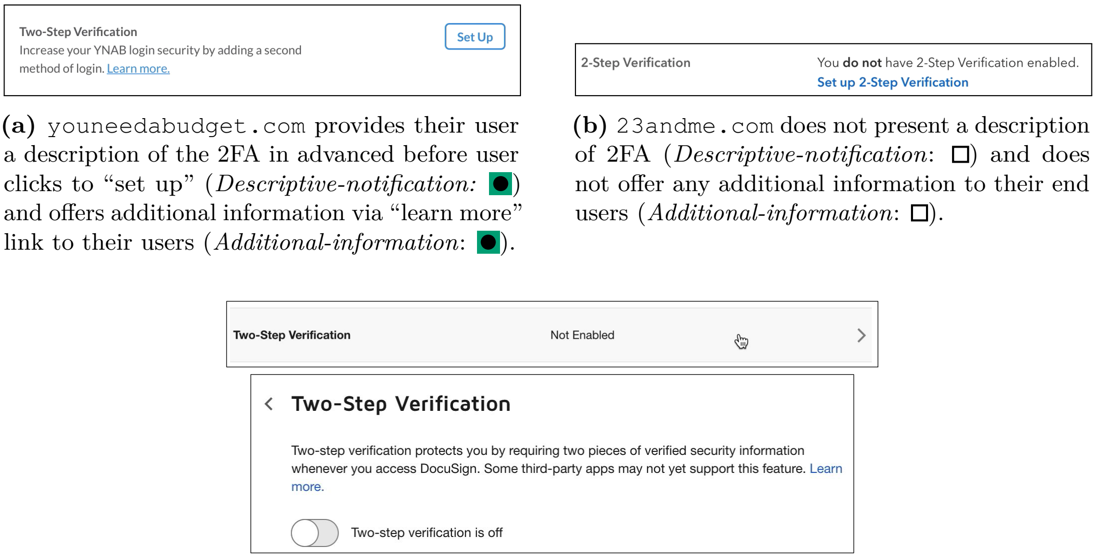
Figure 4.3: Examples of websites that (quasi/not) matched the *Common-Naming-and-Location* factor.

4.6.2 Factors for Education

Descriptive-Notification: The website briefly describes what 2FA is in general or why it is important to users. The description is provided to the user *before* the user clicks to enable 2FA (●), e.g., located together with a notification about 2FA availability or within the settings page; or the description is only provided *after* the user starts the 2FA setup process (◐) at which point the user can still abort the setup. If the website does not present a description of 2FA, we categorize this website as not matching (□).

Additional-Information: The website provides more detailed information through a link (e.g. "learn more") to help users understand 2FA (●). If no such information is provided or the link is broken, the factor does not match (□).

Figure 4.4 illustrates examples for both the *Descriptive-Notification* and *Additional-Information* factors.



Two-Step Verification

Increase your YNAB login security by adding a second method of login. Learn more.

(a) `youneedabudget.com` provides their user a description of the 2FA in advanced before user clicks to "set up" (*Descriptive-notification:* ●) and offers additional information via "learn more" link to their users (*Additional-information*: ●).

2-Step Verification

You **do not** have 2-Step Verification enabled.

(b) `23andme.com` does not present a description of 2FA (*Descriptive-notification*: □) and does not offer any additional information to their end users (*Additional-information*: □).

Two-Step Verification — Not Enabled

Two-Step Verification

Two-step verification protects you by requiring two pieces of verified security information whenever you access DocuSign. Some third-party apps may not yet support this feature. Learn more.

Two-step verification is off

(c) `docusign.com` provides their user a description of the 2FA after user "clicks to enable" (*Descriptive-notification*: ◐) and offers additional information via "learn more" link to their users (*Additional-information*: ●).

Figure 4.4: Examples of websites that (quasi/not) matched *Descriptive-notification* and (not) matched *Additional-information* factors.

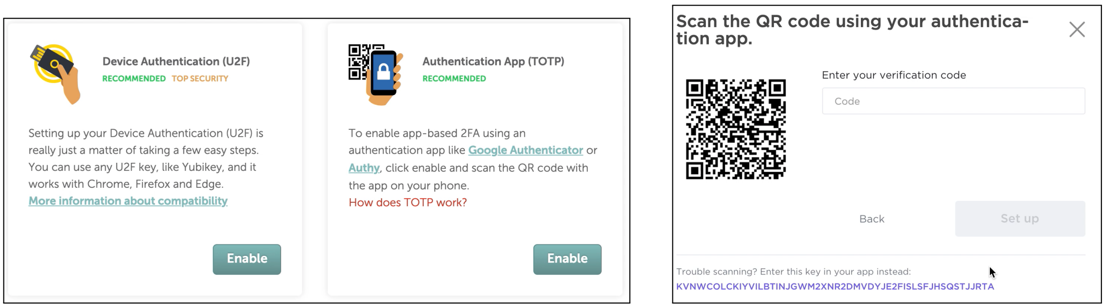
4.6.3 Factors for Setup

Option-Specific-Information: The website provides specific information about all 2FA options it supports (●). For instance, it informs the user that TOTP or Push-notifications require the installation of an app or that WebAuthn requires a hardware authenticator. If the website does not provide this information but directly starts the setup process (e.g., asking users to scan a QR code or to use a security key without further explanation), this factor does not match (□) (see Figure 4.5).

Step-Wise-Instructions: The website gives an overview of the steps involved in setting up a specific 2FA option (e.g., linking a device or app, verifying the link, setting a recovery option) and/or details the instructions for each step for all 2FA options (●). Otherwise, this factor does not match (□) (see Figure 4.6).

Multiselection: The website offers multiple 2FA options (or setting up one method multiple times) and allows the user to set up multiple 2FA options (●), e.g., TOTP and Push-notification or multiple security keys. If the website supports multiple 2FA options but only allows the user to select one non-repeated option, we categorize this as *quasi-multiselection* (◐). This factor does not match (□) if the website only offers a single, one-time configurable option (see Figure 4.7).

Grouped-Setting: The website's user settings present the 2FA options grouped, and users have a single setting location to manage all their 2FA options (●), e.g., all under the same settings tab. If the 2FA settings are split between different sections of the settings, we consider this to be not matching this factor (□). For instance, the management



Device Authentication (U2F)
RECOMMENDED TOP SECURITY

Setting up your Device Authentication (U2F) is really just a matter of taking a few easy steps. You can use any U2F key, like Yubikey, and it works with Chrome, Firefox and Edge.
More information about compatibility

Enable

Authentication App (TOTP)
RECOMMENDED

To enable app-based 2FA using an authentication app like Google Authenticator or Authy, click enable and scan the QR code with the app on your phone.
How does TOTP work?

Enable

Scan the QR code using your authentication app.

Enter your verification code
Code

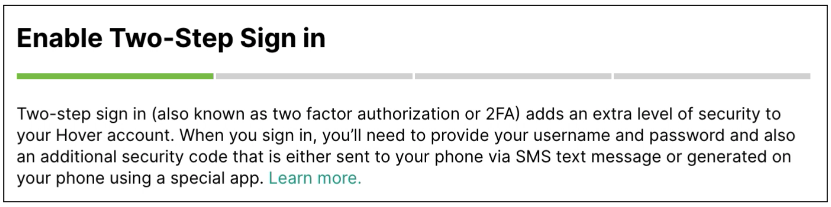
Back Set up

Trouble scanning? Enter this key in your app instead:
KVNWCOLCKIYVILBTINJGWM2XNR2DMVDYJE2FISLSFJHSQSTJJRTA

(a) `namecheap.com` (●) is an example of a website that provides specific information about the 2FA options that it supports.

(b) `clickup.com` (□) does not explain the setup of TOTP to its users but immediately prompts the user to make use of the authenticator app.

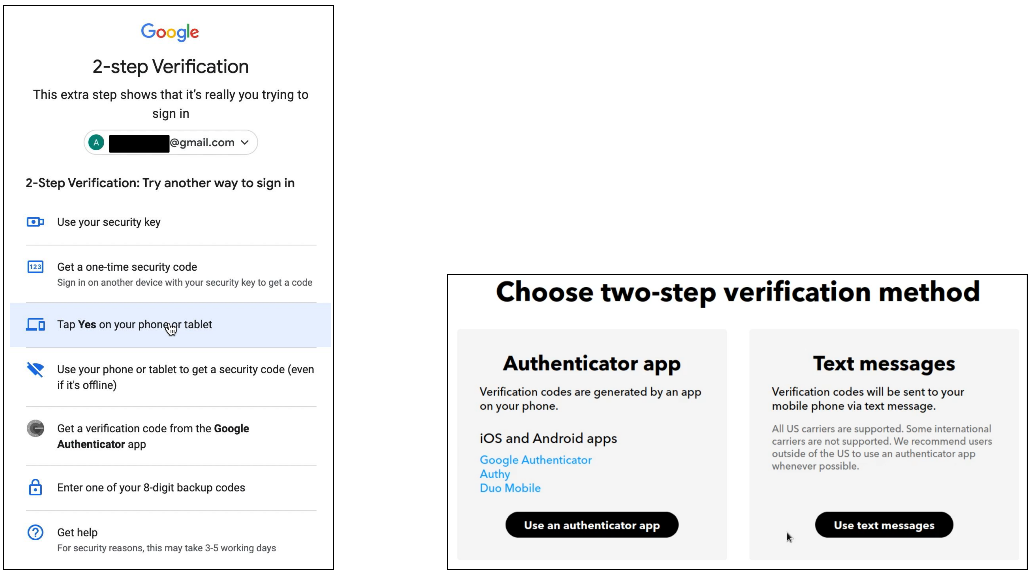
Figure 4.5: Examples that (not) matched the *Option-Specific-Information* factor.



Enable Two-Step Sign in

Two-step sign in (also known as two factor authorization or 2FA) adds an extra level of security to your Hover account. When you sign in, you'll need to provide your username and password and also an additional security code that is either sent to your phone via SMS text message or generated on your phone using a special app. Learn more.

Figure 4.6: `hover.com` (●) is an example that matched the *Stepwise-Instruction* factor. It gives an overview of the involved steps of setting up 2FA as a progress bar.



Google
2-step Verification
This extra step shows that it's really you trying to sign in
A [redacted]@gmail.com

2-Step Verification: Try another way to sign in

- Use your security key
- Get a one-time security code
  Sign in on another device with your security key to get a code
- Tap **Yes** on your phone or tablet
- Use your phone or tablet to get a security code (even if it's offline)
- Get a verification code from the **Google Authenticator** app
- Enter one of your 8-digit backup codes
- Get help
  For security reasons, this may take 3-5 working days

Choose two-step verification method

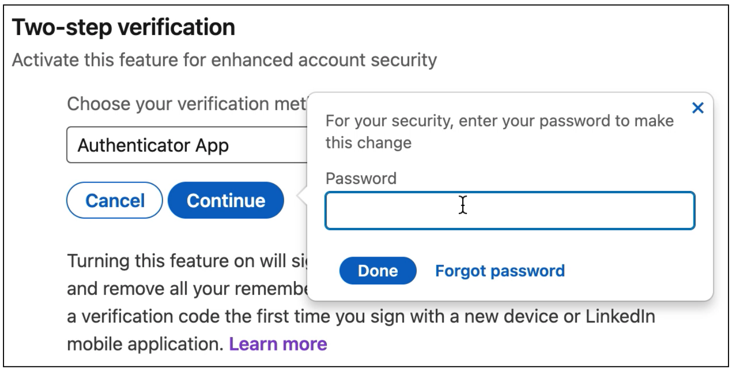
Authenticator app
Verification codes are generated by an app on your phone.
iOS and Android apps
Google Authenticator
Authy
Duo Mobile
Use an authenticator app

Text messages
Verification codes will be sent to your mobile phone via text message.
All US carriers are supported. Some international carriers are not supported. We recommend users outside of the US to use an authenticator app whenever possible.
Use text messages

(a) `google.com` (●) offers multiple 2FA options and allows the user to setup and activate multiple 2FA options at the same time.

(b) `ifttt.com` (◐) offers multiple 2FA options but allows the user to only select one active option at the same time.

Figure 4.7: Examples of websites that (quasi) matched the *Multiselection* factor.



Two-step verification

Activate this feature for enhanced account security

Choose your verification met

Authenticator App

Cancel Continue

Turning this feature on will si
and remove all your rememb
a verification code the first time you sign with a new device or LinkedIn
mobile application. Learn more

×

For your security, enter your password to make this change

Password

Done Forgot password

Figure 4.8: `linkedin.com` is an example for a website that asks users to verify their identity before being able to setup 2FA (*Settings-Changed-Verification*: ●).

of security keys is organizationally separated from managing other 2FA options and, hence, might not be obvious to users. This factor depends on *Multiselection* being (quasi-)matched.

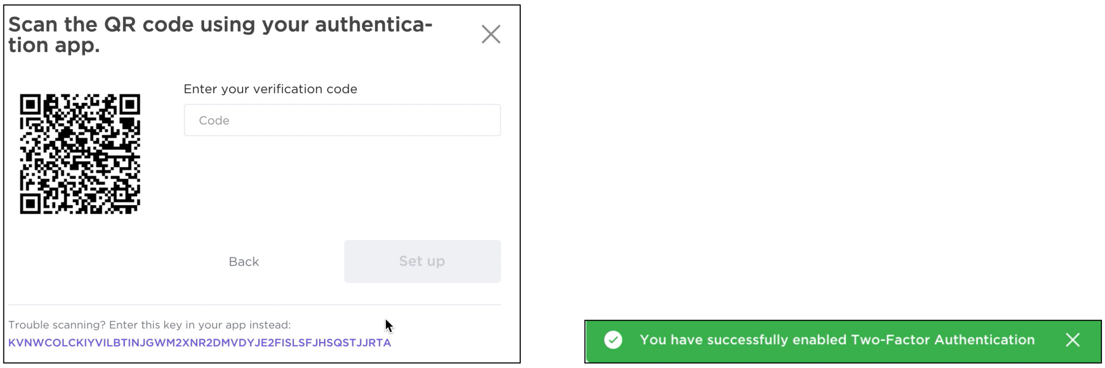
No-Enforced-Options: The website immediately presents all supported 2FA options to the user and allows them to choose their options themselves (●). If the website mandates the setup of specific 2FA options before the user can set up other options, we consider this not to match this factor (□). For example, the user must configure SMS-based 2FA before having the possibility to configure TOTP or WebAuthn options. This factor depends on *Multiselection* being (quasi-)matched.

Selectable-Primary-Option: If the website allows the configuration of multiple 2FA options and allows the user to select a primary option, which is the first option requested during login before falling back to other configured options (or recovery), we consider this a match (●). If the website does not support setting a user-selected primary 2FA option, we consider this not matching (□). This factor depends on *Multiselection* being matched.

Settings-Changed-Verification: The website requires the user to verify their identity before being able to change the 2FA settings (●). Otherwise, this factor does not match (□) (see Figure 4.8).

Settings-Changed-Notification: The website notifies the user about the changed 2FA settings via an out-of-band channel, e.g., by email or push notification (●). If there is no notification, this website does not match this factor (□).

Confirm-Successful-Setup: The website requires the user to confirm the 2FA authentication to complete the setup successfully and provides clear messaging about the successful setup for all options (●). For example, the user must enter the current TOTP or confirm a push notification to complete the setup, and the website shows a highlighted message in the settings. If the messaging is missing, but confirmation is required, we consider this as *quasi-confirm-successful-setup* (◐). If the website does not require confirmation (for all options), this website does not match this factor (□) (see Figure 4.9).



Scan the QR code using your authentication app.

Enter your verification code

Trouble scanning? Enter this key in your app instead:

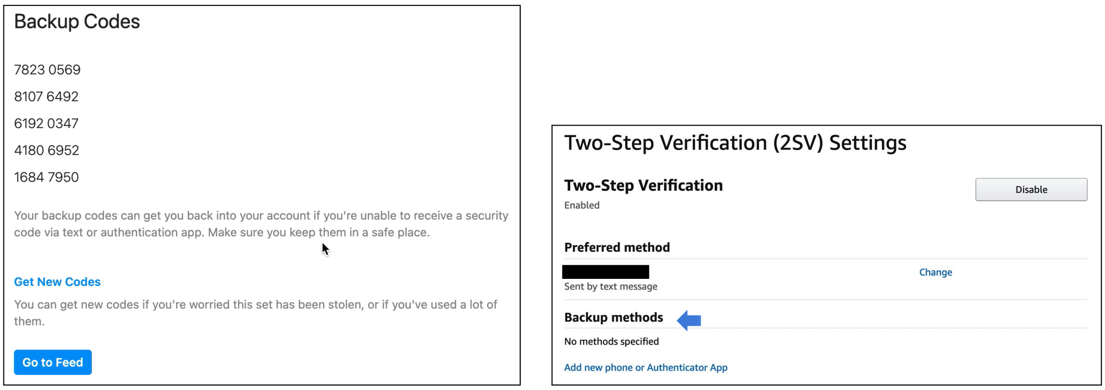
KVNWCOLCKIYVILBTINJGWM2XNR2DMVDYJE2FISLSFJHSQSTJJRTA

You have successfully enabled Two-Factor Authentication

(a) `clickup.com` asks users to confirm the successful setup of TOTP by entering the current code.

(b) `clickup.com` shows a highlighted message to inform the user about the successful activation of 2FA.

Figure 4.9: `clickup.com` as an example for websites that matched the *Confirm-Successful-Setup* (●) factor.



Backup Codes

7823 0569
8107 6492
6192 0347
4180 6952
1684 7950

Your backup codes can get you back into your account if you're unable to receive a security code via text or authentication app. Make sure you keep them in a safe place.

Get New Codes

You can get new codes if you're worried this set has been stolen, or if you've used a lot of them.

Two-Step Verification (2SV) Settings

Two-Step Verification
Enabled

Preferred method
[redacted]
Sent by text message

Backup methods
No methods specified

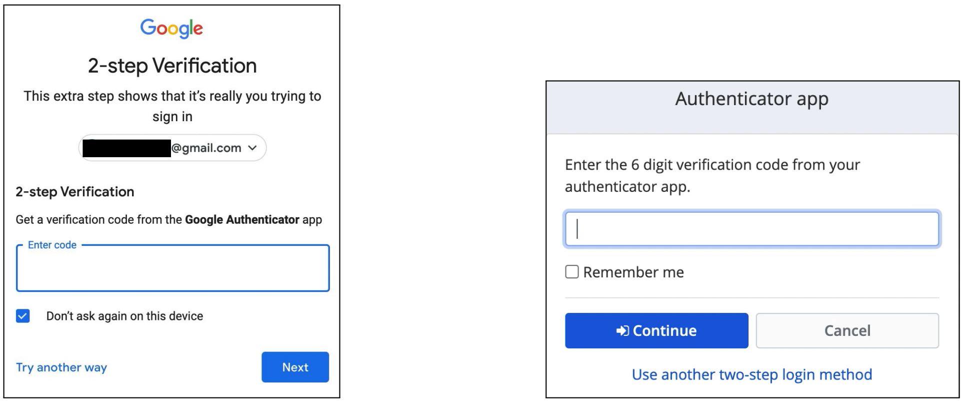
(a) `instagram.com` (●) offers a dedicated recovery option and explains to the user why it is important to enable recovery options.

(b) `clickup.com` (◐) offers dedicated recovery options but does not provide any explanation to the user why it is important to enable recovery options.

Figure 4.10: Examples that (quasi) matched the *Informed-2FA-Recovery-Options* factor.

Informed-2FA-Recovery-Options: The website offers dedicated recovery options (such as one-time codes or asking to set up multiple 2FA options) and explains to the user why configuring dedicated 2FA recovery options is important for preparing for cases where the default 2FA options are not available, e.g., to prevent account lockout due to a lost or broken authentication device (●). If the website offers such recovery options but does not explicitly inform the user about their benefits and importance, we consider this as *quasi-informed-recovery-options* (◐). If the website does not offer explicit 2FA recovery options (e.g., it relies on a general account recovery or customer support), we consider this as not matching this factor (□) (see Figure 4.10).

Enforced-2FA-Recovery-Setup: Setting up recovery options is a mandatory step in setting up 2FA for this website (●), and the user cannot finish or continue setting up 2FA unless they set up the recovery option first. For example, the user has to confirm



Google

2-step Verification

This extra step shows that it's really you trying to sign in

[redacted]@gmail.com

2-step Verification

Get a verification code from the **Google Authenticator** app

Enter code
☑ Don't ask again on this device
Try another way
Next

Authenticator app

Enter the 6 digit verification code from your authenticator app.

□ Remember me
Continue
Cancel
Use another two-step login method

(a) `google.com` (◐) puts the device remembrance at the discretion of the user and states it as opt-*out*.

(b) `meistertask.com` (◐) states the device remembrance as opt-*in*.

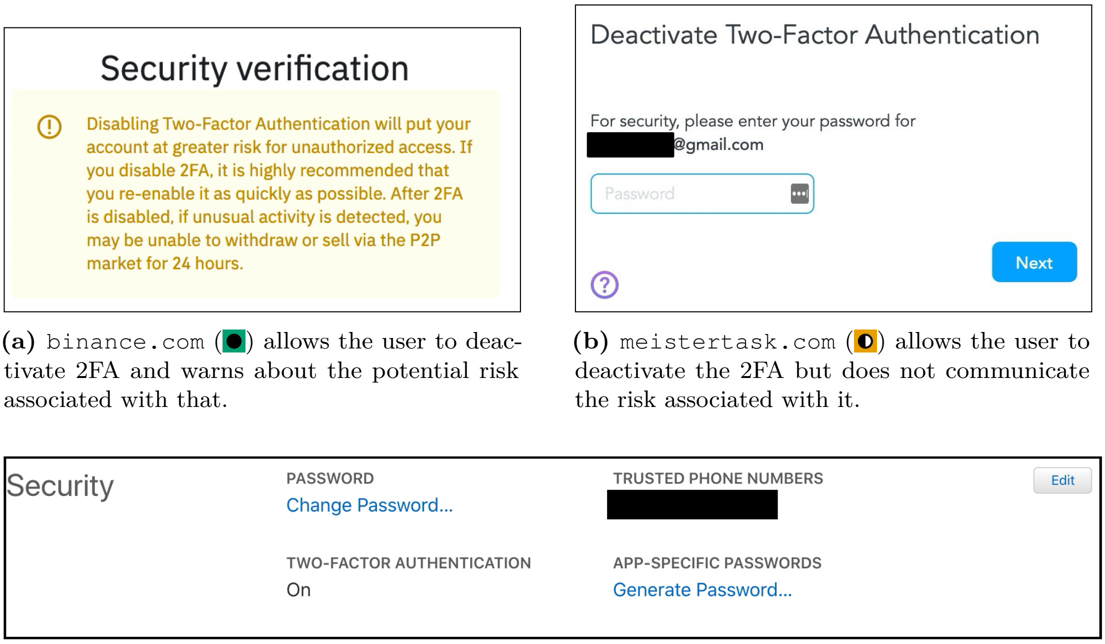
Figure 4.11: Examples that (quasi) matched the *Device-remembrance* factor.

that they printed one-time backup codes to finish the 2FA setup or the website enforces setting up multiple 2FA options with a clear hint at account recovery. If setting up recovery options is not mandatory, but the website nudges users or strongly recommends them to set up a recovery option, we consider this *quasi-mandatory-recovery-setup* (◐). If setting up dedicated recovery options is at the user's discretion (without nudging or recommending), we consider this factor not matching (□). This factor depends on *Informed-2FA-recovery-options* being (quasi-)matched.

4.6.4 Factors for Usage

Device-Remembrance: The website offers a device remembrance during login, such that the user does not have to use 2FA on subsequent logins on the same device (e.g., "remember this device" checkbox). If the website automatically sets device remembrance without involving the user, e.g., during the first login after 2FA setup or during 2FA setup, we categorize this as ●. If device remembrance is at the discretion of the user and is stated as opt-*out* (e.g., an unchecked checkbox described as "ask me again on this device" or a pre-ticked checkbox "trust this device"), we categorize this as ◐. If device remembrance is stated as opt-*in* (e.g., "trust this device" checkbox that was not pre-checked), we categorize this as ◐. If device remembrance is not offered, we categorize as □ (see Figure 4.11).

No-Preselected-Option: If the website supports more than one active 2FA option at a time and no primary method is set (or could be set), how does the website present the configured 2FA options to their end users: the website shows all configured 2FA options at the same time during login (●), e.g., as a drop-down list. Alternatively, the website selected the primary option based on internal metrics (□), e.g., a security policy or the user's usage history. This preselection is usually intransparent to the user. This factor depends on *Multiselection* being matched and *Selectable-primary-option* not being matched.



Security verification

Disabling Two-Factor Authentication will put your account at greater risk for unauthorized access. If you disable 2FA, it is highly recommended that you re-enable it as quickly as possible. After 2FA is disabled, if unusual activity is detected, you may be unable to withdraw or sell via the P2P market for 24 hours.

(a) `binance.com` (●) allows the user to deactivate 2FA and warns about the potential risk associated with that.

Deactivate Two-Factor Authentication

For security, please enter your password for [redacted]@gmail.com

(b) `meistertask.com` (◐) allows the user to deactivate the 2FA but does not communicate the risk associated with it.

Security

PASSWORD
Change Password...

TRUSTED PHONE NUMBERS
[redacted]

TWO-FACTOR AUTHENTICATION
On

APP-SPECIFIC PASSWORDS
Generate Password...

(c) `icloud.com` (□) does not allow the user to deactivate the 2FA.

Figure 4.12: Examples that (quasi/not) matched the *Informed-deactivation* factor.

4.6.5 Factors for Deactivation

Informed-Deactivation: The website allows the user to deactivate 2FA options and also explains to the user the potential risks associated with this (●) or does not provide any explanation or warning (◐). If the website does not allow the user to deactivate two-factor authentication, we consider this a mismatch for this factor (□).

Deactivation-Verification: The website requires the user to verify their identity before being able to deactivate a 2FA option (●). If a 2FA option can be disabled by the user without further authorization, we consider this factor to not match the website (□). This factor depends on *Informed-deactivation* being (quasi-)matched.

Deactivation-Notification: The website notifies the user about the deactivated 2FA option via an out-of-band channel, e.g., by email (●). If there is no notification, this website does not match this factor (□). This factor depends on *Informed-deactivation* being (quasi-)matched (see Figure 4.12).

Communicate-Successful-Deactivation: The website communicates successful deactivation to the user as part of its user interface (●), e.g., highlighted message or pop-up. Otherwise, we consider this website not to match this factor (□). This factor depends on *Informed-deactivation* being (quasi-)matched.

			Discovery			Edu.		Setup											Usage		Deactiv.			
Website	**Subcluster**	**Category**	Promotion	Non-optional	Common-Naming-and-Location	Descriptive-notification	Additional-information	Option-specific-information	Stepwise-instructions	Multiselection	Grouped-setting	No-enforced-option	Selectable-primary-option	Settings-changed-verification	Settings-changed-notification	Confirm-successful-setup	Informed-2FA-recovery-options	Enforced-2FA-recovery-setup	Device-remembrance	No-preselected-option	Informed-deactivation	Deactivation-verification	Deactivation-notification	Communicate-successful-deact.
Cluster 1 ($n = 30$)																								
airvpn.org	1	VPN Providers			◐	◐		●			■	■	■	●		●		■		■	◐		●	
booking.com	1	Hotels/Accom.			●			●			■	■	■			◐	◐			■	◐			
clickup.com	1	Task Management			●	●				◐	●		■			●		■		■	◐			
clio.com	1	Legal	◐		◐	●					■	■	■				●	◐		■	◐			●
digicert.com	1	Security			●	●		●	●		■	■	■					■	◐	■	◐			
instagram.com	1	Social			●	◐				◐	●		■			◐	●	◐		■	◐			
laravel.com	1	Cloud Computing			●						■	■	■				●	◐		■	◐			
mega.io	1	Backup and Sync	◐		●	●					■	■	■			●	●	◐		■	◐		●	
orcid.org	1	Identity Management			●	◐	●	●			■	■	■			◐	●	◐		■	◐			
runsignup.com	1	Health			◐	◐		●			■	■	■			●	◐			■	◐			●
teamviewer.com	1	Remote Access	●		◐	●	●	●			■	■	■			●	●	◐		■	◐		●	●
toodledo.com	1	Task Management			◐	●		●			■	■	■			●		■		■	◐			●
1password.com	2	Identity Management			◐	●	●			●	●					●		■	●		◐	●	●	●
airtable.com	2	Task Management			●	◐	●	●		●	●		●			●	●			■	◐	●		
arlo.com	2	IoT	●	●	◐	●	●	●		●	●	●	●			◐	◐	●		■	◐			
easydns.com	2	Domains			●					◐	●	●	■			●	●	◐		■	◐	●	●	
gitlab.com	2	Developer			●	●		●		●	●						●	◐			◐	●		●
roboform.com	2	Identity Management			◐	●				◐	●	●	■		●	●		■	●	■	◐	●		●
bitdefender.com	3	Security			●	●				◐	●	●	■			●	●	◐	◐	■	●		●	
blockchain.info	3	Cryptocurrencies			●	●				◐	●	●	■					■	●	■	◐		●	●
coned.com	3	Utilities	●	●	●	●				◐	●	●	■					■		■		■	■	■
facebook.com	3	Social			●	●		●		●	●	●				◐	●		●		◐			●
hover.com	3	Domains	◐		◐	◐		●	●	◐	●	●	■			●	●	◐		■	◐		●	
join.me	3	Remote Access			●	●	●	●	●	◐	●	●	■			●	●	●	●	■	◐			●
jottacloud.com	3	Backup and Sync			●	●				◐	●	●	■			●	●	◐		■	◐		●	
kraken.com	3	Cryptocurrencies	●	●	●	●				◐	●	●	■		●	◐	◐			■		■	■	■
logmein.com	3	Remote Access			●	●	●	●	●	◐	●	●	■				●	●	◐	■	◐			●
mailchimp.com	3	Communication			●	●				◐		●	■			●	●	◐	◑	■	◐			●
namecheap.com	3	Domains	◐		●	●		●		◐	●	●	■	●		●	●	◐		■	●		●	●
xero.com	3	Finance			●	●	●	●	●	◐	●	●	■				●	●	◐	■	◐		●	●
Cluster 2 ($n = 29$)																								
bitwarden.com	1	Identity Management			◐	●		●	●	●	●			●		●	●	◐	◐		◐		●	●
blizzard.com	1	Gaming			◐	●					■	■	■	●	●			■		■	●	●	●	●
callcentric.com	1	Utilities	●			●					■	■	■	■	●			■	◐	■	◐	●		
clubhouse.io	1	Task Management			●	●					■	■	■	●	●	●	●	◐		■	◐	●	●	●
icloud.com	1	Backup and Sync	●	●	●						■	■	■	●		◐		■	◐	■		■	■	■
keepersecurity.com	1	Identity Management	◐		●					●					●	●	●	◐	◐		◐	●		●
kickstarter.com	1	Crowdfunding			●	◐	●			●				●	●	●		■	◐		◐	●		●
realvnc.com	1	Remote Access	●		●	●					■	■	■	●	●	◐	●	◐		■	●	●	●	
reddit.com	1	Social			●	●					■	■	■	●	●	●	●	◐		■	◐	●	●	
roblox.com	1	Gaming			●	●					■	■	■	●	●			■	◐	■	◐	●	●	
synology.com	1	Backup and Sync	●		●	●				●	●			●		●	●	◐	◑		◐		●	
virustotal.com	1	Security			●	●					■	■	■	●		◐	●	◐	◐	■	◐		●	
adobe.com	2	Other			●	●				●	●	●		●	●	●		■	◑		◐	●	●	●
backblaze.com	2	Backup and Sync	●		◐					◐	●	●	■		●	●	●	◐	◐	■	◐	●		●
bybit.com	2	Cryptocurrencies			●					●	●	●		●	●	●		■			◐	●	●	●
docusign.com	2	Legal			●	◐	●		●	●	●	●		●	●	●		■	◐		◐	●	●	●
dropbox.com	2	Backup and Sync			●	◐	●	●		●	●			●	●	●	●	◐	◐		◐	●	●	●
ea.com	2	Gaming			◐	●				◐	●	●	■	●	●	◐	●	◐	●	■	●	●	●	
evernote.com	2	Backup and Sync			●	◐		●		●	●		●		●	●	●	◐	◐	■	◐	●		

continued on next page

			Discovery			Edu.		Setup											Usage		Deactiv.			
Website	**Subcluster**	**Category**	Promotion	Non-optional	Common-Naming-and-Location	Descriptive-notification	Additional-information	Option-specific-information	Stepwise-instructions	Multiselection	Grouped-setting	No-enforced-option	Selectable-primary-option	Settings-changed-verification	Settings-changed-notification	Confirm-successful-setup	Informed-2FA-recovery-options	Enforced-2FA-recovery-setup	Device-remembrance	No-preselected-option	Informed-deactivation	Deactivation-verification	Deactivation-notification	Communicate-successful-deact.
Cluster 2 (continued)																								
lastpass.com	2	Identity Management	◐		◐	●	●			●	●		●	●		●	◐		◐	■	◐	●	●	
norton.com	2	Security	●		●	●	●			●	●	●	●		●		●	◐	◐	■	◐	●		
stripe.com	2	Payments			◐	●				●	●				●	●	●	◐			◐	●	●	●
twitch.tv	2	Entertainment			●	●				●	●				●	●	●	●	◐		●	●		●
unity.com	2	Developer			●			●		●	●	●	●	●	●	●	◐		●	■	◐	●		●
vk.com	2	Social	●		●	●				●	●			●	●	●	●	◐	◑		◐	●	●	
zoho.com	2	Email			●	●				●	●	●	●		●	●	●	◐	◐	■	●	●		●
zoom.us	2	Communication	◐		●					●	●	●		●	●	●	●	◐			◐	●	●	
cloudflare.com	3	Security			●	●	●			●	●	●		●			●	◐			◐		●	●
tumblr.com	3	Social			●	●	●			●	●	●		●		◐	●				●		●	
Cluster 3 ($n = 4$)																								
23andme.com	1	Health			●				●		■	■	■	●		●		■		■	◐		●	●
discord.com	1	Communication			●	●			●	●				●		●	●	◐			◐		●	
opera.com	1	Other	●		◐	●			●		■	■	■			●	●	◐		■	◐			●
gandi.net	3	Domains	●		●	◐			●	●		●		●		●	●	◐			◐			●
Cluster 4 ($n = 9$)																								
atlassian.com	1	Developer			●	●	●	●	●		■	■	■	●	●	●	●	◐		■	●	●	●	●
basecamp.com	2	Communication			●	●		●	●	●	●				●	●	●	◐			●	●		●
binance.com	2	Cryptocurrencies	●	●	●	●			●	●	●			●	●	●	●	◐			●	●	●	●
bitfinex.com	2	Cryptocurrencies	●	●	●	●		●		●	●	●		●	●	●		■			●	●	●	●
digitalocean.com	2	Cloud Computing			●	●			●	◐	●	●	■		●	◐	●	◐		■	●	●		●
google.com	2	Email			●	●	●	●	●	●	●			●	●		●	◐	◑		●	●	●	●
newegg.com	2	Retail			●	●		●	●	◐	●	●	■		●	●	●	◐	◐	■	●	●		
epicgames.com	3	Gaming			●	●	●	●	●	●	●	●	●		●	●	●	◐	◑	■	●		●	●
id.me	3	Identity Management			◐	●		●	●	●	●	●			●			■		●		■	■	■
Cluster 5 ($n = 8$)																								
github.com	1	Developer			●	●	●	●		●					●	●	●	◐			◐	●		
meistertask.com	1	Task Management	◐		●	●	●	●			■	■	■	●	●	●	●	◐		■	◐	●	●	
youneedabudget.cc	1	Finance			●	●	●	●	●		■	■	■		●	●	●	◐		■	◐	●	●	●
ifttt.com	2	IoT			●	●	●	●		◐	●	●	■	●	●	◐	●	◐		■	◐	●	●	
playstation.com	2	Gaming			●	◐	●	●		◐	●	●	■		●	●	●	◐		■	●	●		●
twitter.com	2	Social			●	●	●	●		●	●	●		●	●	●	●	◐			◐	●		
etsy.com	3	Retail			●	●	●	●		◐	●	●	■		●	●	●	◐	◐	■	◐			
va.gov	3	Government	●		●	●	●	●	●	◐	●	●	■		●	●	●	◐		■		■	■	■
Cluster 6 ($n = 5$)																								
amazon.com	2	Retail			●	◐		●	●	◐	●	●	■	●	●	●	◐		◐	■	◐	●		●
ebay.com	2	Retail			●	●		●	●	◐	●	●	■	●	●	●		■		■	◐	●	●	●
linkedin.com	2	Social			●	●	●		●	◐	●	●	■	●	●	●	◐		◑	■	◐	●	●	
paypal.com	2	Payments			●	●			●	◐	●	●	■		●	●	◐		◐	■	●	●		
yahoo.com	2	Email			●	●		●		◐	●	●	■	●	●	●	◐	●	◑	■	◐	●		●

● = factor matches; □ = factor does not match; ■ = does not apply

◐ = factor quasi matches (factors with 3-point and 4-point scales); ◑ = factor quasi matches (factors with 4-point scales)

Gray column headers indicate conditional factors dependent on other factors to be applicable

Table 4.1: Comparison of popular websites based on the factors introduced in Section 4.6 and clustering described in Section 4.7.

Name of 2FA in website settings	
Two-Factor Authentication (2FA)	42 (49.41%)
Two-Step Verification (2SV)	24 (28.24%)
Other	12 (14.12%)
Multi-Factor Authentication (MFA)	4 (4.71%)
Two-Step Authentication (2SA)	3 (3.53%)
Location of 2FA settings	
Security / Account	78 (91.76%)
Other	7 (8.24%)
Focus of 2FA description	
Security	69 (92.00%)
Device	6 (8.00%)
Description of device remembrance	
Remember	16 (51.61%)
Trust	9 (29.03%)
Skip	4 (12.90%)
Other	2 (6.45%)

Table 4.2: Naming, location, 2FA descriptions, and description of remembrance.

4.7 Results

We first provide an overview of the collected data (Section 4.7.1), followed by exploratory data analysis of the comparison factors (Section 4.7.2). Lastly, we discuss the results of qualitative data analysis of our observations (Section 4.7.3).

4.7.1 Overview of Website Data

Table 4.1 summarizes how each of the 85 websites in our data set matches the 22 comparison factors that we identified. We will explore these data further in the following sections. Table 4.2 summarizes the naming and location of the 2FA settings, the type of 2FA description, and the forms of device remembrance. Table 4.3 provides the codebook for device remembrance descriptions. Table C.3 in Appendix C.2 gives further details per website, including Tranco [125, 184] rank and website category according to 2fa.directory [3, 2]. In summary, we found that 73 (86%) of the websites use a combination of “two-factor”/“two-step”/“multiple-factor” with “authentication”/“verification” for the naming, and on 78 (92%) websites, the 2FA settings are located in the security settings of the account settings under similar paths (e.g., “Security,” “Login security,” or “Authentication”). We considered those names and locations the common naming and location during our evaluation of the *Common-Naming-and-Location* factor. Of the 75 websites that describe 2FA in their settings, 69 (92%) describe 2FA in the form of “an additional layer of security,” while 6 websites describe the 2FA mechanism with a focus on the user device (e.g., “we ask for additional authentication when logging in from a device that we do not know”). Only 31 websites in our data set offered a device remembrance feature, and half of those (16; 52%) describe this feature in terms

Code	Examples
Remember	Remember verification for this computer
	Recognize this device in the future
	Do not require OTP on this browser
	Skip two-factor authentication on this device
	Save browser
	Do not ask again on this device
	Remember this device
	Remember this computer for {duration}
	Do not ask for code on this device
Trust	Trust this device (opt-in)
	Trust this device (opt-out)
	Do not trust this device (opt-out)
	Do not trust this device (opt-in)
	Trust this device for {duration}
	Untrust this device
Skip	Require code to login for {duration}
	We won't ask for the next {duration}
	Skip this for {duration}
Other	Stay signed/logged in

opt-out: checkbox is pre-checked; *opt-in*: checkbox is not pre-checked
duration: a number of days, weeks, or logins

Table 4.3: Codebook for device remembrance.

of remembering the device or client (e.g., "Do not require OTP on this browser" or "Do not ask again on this device"). Almost a third (9; 29%) describe it in terms of trust (e.g., "Trust this device for {duration}"), and only four websites (13%) phrase it as skipping the additional step (e.g., "We won't ask for the next {duration}"). We also noticed that websites have a mixed strategy for phrasing the user's choice (i.e., opt-*in* versus opt-*out*), which we encoded in our factor *Device-remembrance* in Table 4.1 (i.e., ◐ vs. ◑).

4.7.2 Exploratory Data Analysis

Our factors allow us to compare the 2FA user journeys of different websites. We first explore our collected data (in Table 4.1) through similarity analysis and clustering to gain insight into the overall consistency of those journeys on different websites and to identify potential clusters of websites that follow similar design patterns for their 2FA user experience.

4.7.2.1 Website similarity and factor consistency

To get a general impression of how similar the journeys are on the 85 websites, we compared them pairwise. Since our comparison factors are feature vectors of nominal (i.e., categorical) variables for each website, there is no intrinsic ordering and no equal space between variable values to measure the distance between values. We used the Hamming distance between pairs of websites as a measure of similarity. Since our variables have

	Comparison Factor	$H(X)$	Max ent.
Two-point scale	Non-optional	0.37	1.0
	Additional-information	0.90	
	Option-specific-information	0.99	
	Stepwise-instructions	0.87	
	Settings-changed-verification	0.99	
	Settings-changed-notification	1.00	
Three-point scale	Promotion	1.12	1.57
	Descriptive-notification	1.11	
	Multiselection	1.57	
	Confirm-successful-setup	1.24	
	Informed-2FA-recovery-options	1.26	
	Informed-deactivation	1.05	
Four-point scale	Common-Naming-and-Location	1.00	2.0
	Device-remembrance	1.60	

Table 4.4: Shannon entropy of each non-conditional factor.

only values between 2–4, Hamming distance (i.e., “overlap without weights”) is the most efficient measure of similarity for our data to obtain an overall impression of consistency between websites instead of measures that consider the number and/or frequency of values per variable, such as (Inverse) Occurrence Frequency, Goodall [83], or Eskin et al. [62]. To avoid artificial inflation of similarity from unfulfilled conditional factors, we calculate the Hamming distance only for the 14 non-conditional factors. We find that the average website in our data set differs in 6–7 of those 14 factors from the other websites, indicating that from a bird’s-eye view, the user journeys are not very consistent across those websites. Further details about the frequency distribution of the pairwise Hamming distances are provided in Appendix C.3.

Furthermore, we measured the consistency of individual, non-conditional factors across all websites using Shannon entropy. High entropy means high inconsistency, whereas low entropy implies high consistency. The results are summarized in Table 4.4. Since some factors can also *quasi*-match (◐/◑) and, thus, have a different maximum entropy from binary (two-point scale) factors with only ● and □, we distinguish between the point scales for each factor. The maximum possible entropy for each scale is indicated in column *Max ent.* Noticeable outliers with high entropy, i.e., low consistency, are *Multiselection* and most of the two-scale factors, which are close to the highest possible entropy. For example, *Multiselection* is almost evenly split (34×●, 28×◐, 23×□). In contrast, *Non-optional* is very consistent (6×●, 79×□) and *Common-Naming-and-Location* shows a strong tendency (67×●, 6×◑, 11×◐, 1×□). In summary, we found that none of the factors exhibit high consistency, except for *Non-optional* 2FA and *Common-Naming-and-Location.*

4.7.2.2 Factor clusters

Since our data do *not* indicate a “global consistency,” we explore further whether there exist clusters of websites that have close similarities to each other but are more dissimilar from others. We applied a two-stage clustering process: first, we cluster websites based on their non-conditional factors (*inter-cluster*) and, additionally, assign each website to a subcluster based on the conditional factors (*intra-cluster*). Our intention for this two-stage process was that inter-clusters based on non-conditional factors provide the primary view of the different strategies for the 2FA user journeys across all websites, while additional intra-clusters based on conditional factors could support a more differentiated discussion of the overall strategies. Since our comparison factors are nominal variables, we apply k-modes [28] clustering in both stages. For inter-clustering, Silhouette testing [167] indicated that 2, 5, or 6 clusters fit the data best, and we decided on 6 clusters due to the best descriptive performance of those clusters. For the intra-clustering of the conditional factors, we found 3 clusters to best describe the data. The result of the final clustering is noted in Table 4.1. Figure 4.14 provides less noisy views of the cluster structures.

When comparing the characteristics of the *inter-clusters*, we find three aspects that differentiate the clusters the most: how they inform and instruct their users, how they offer support for multiple 2FA options, and whether they offer device remembrance. In terms of informing and instructing users, the six clusters can be combined into two larger clusters. Websites in ***Clusters 1, 2*** and ***3*** generally do not verify or notify about changes in 2FA settings (except for ***Cluster 2***), omit additional information, do not give step-wise instructions (with the exception of ***Cluster 3***), and often do not provide specific information about 2FA options. In contrast, ***Clusters 4, 5*** and ***6*** provide this information and instructions more regularly and, in addition, the websites in ***Cluster 4*** warn users about the deactivation of 2FA. Alternatively, the six *inter-clusters* could be combined into two groups based on their strategy to support multiple 2FA options. Here, websites in ***Clusters 1, 5*** and ***6*** usually allow only one option to be activated simultaneously, although they usually offer multiple options. In contrast, websites in ***Clusters 2, 3*** and ***4***, when supporting multiple 2FA options, usually allow users to choose between multiple activated 2FA options for login. Lastly, regarding device remembrance, the websites in ***Clusters 1, 3, 4*** and ***5*** have in common that they mostly do not offer device remembrance for future logins. ***Clusters 2*** and ***6*** usually offer this.

Regarding *intra-clusters*, ***Subcluster 1*** websites do not usually provide a selection of multiple 2FA options, and when they do, they enforce certain 2FA options. The websites in ***Subclusters 2*** and ***3*** support multiple 2FA options but differ in their strategy to enforce certain 2FA options and verify 2FA deactivation. Websites in ***Subcluster 2*** almost always verify 2FA deactivation, while websites in ***Subcluster 3*** do not enforce certain 2FA options. Unfortunately, ***Clusters 3*** to ***6*** are too small to reliably comment on the relationship between *inter-clusters* and *intra-cluster*.

4.7.2.3 Clusters vs. Website Ranks

We divide the websites into our data set into three roughly equal-sized groups through the 36th and 71st percentiles of the websites' Tranco ranks. Figure 4.15 illustrates the CFD of the Tranco ranking in our data set. Based on this CFD, the first group of websites ($n = 31$) is in the *Top-500* of Tranco, the second group of websites ($n = 29$) ranks between 501 and 4,000 (denoted as *Top-4000*), and the third group ($n = 25$) is the *"long tail"* with a rank greater than 4,000. Since we initially selected the most

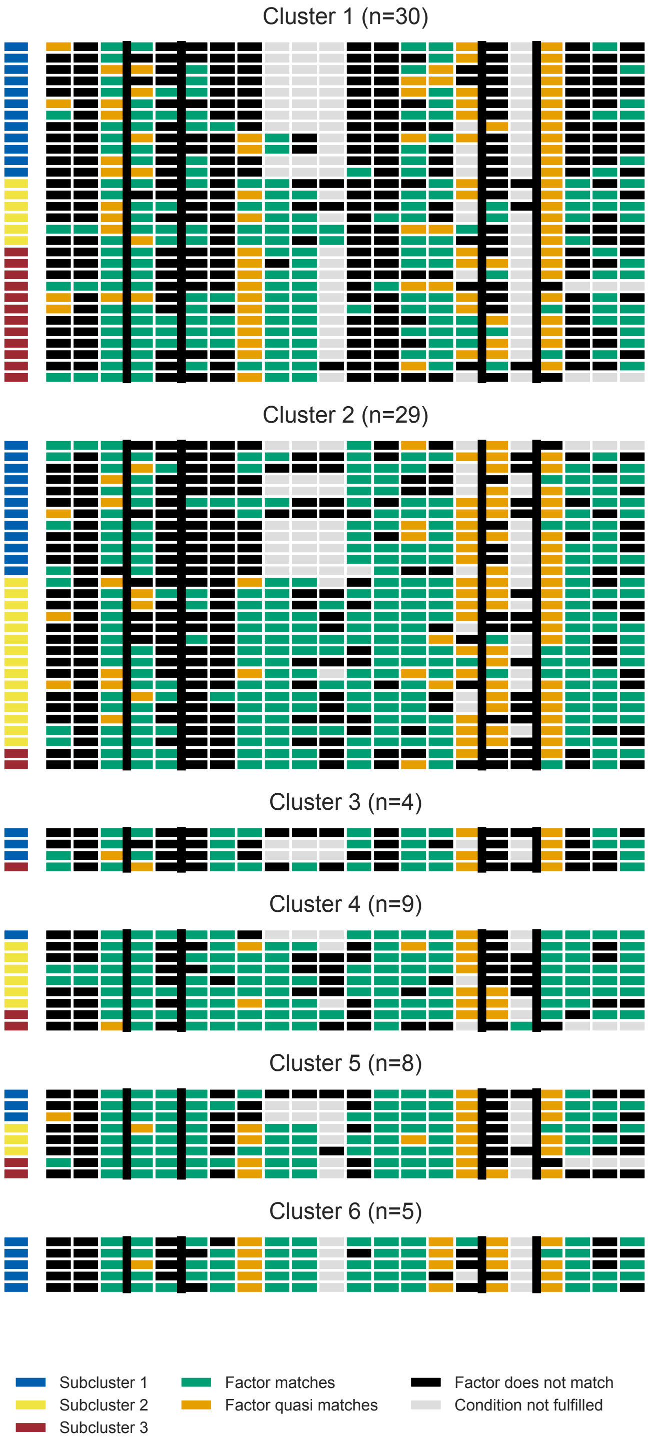
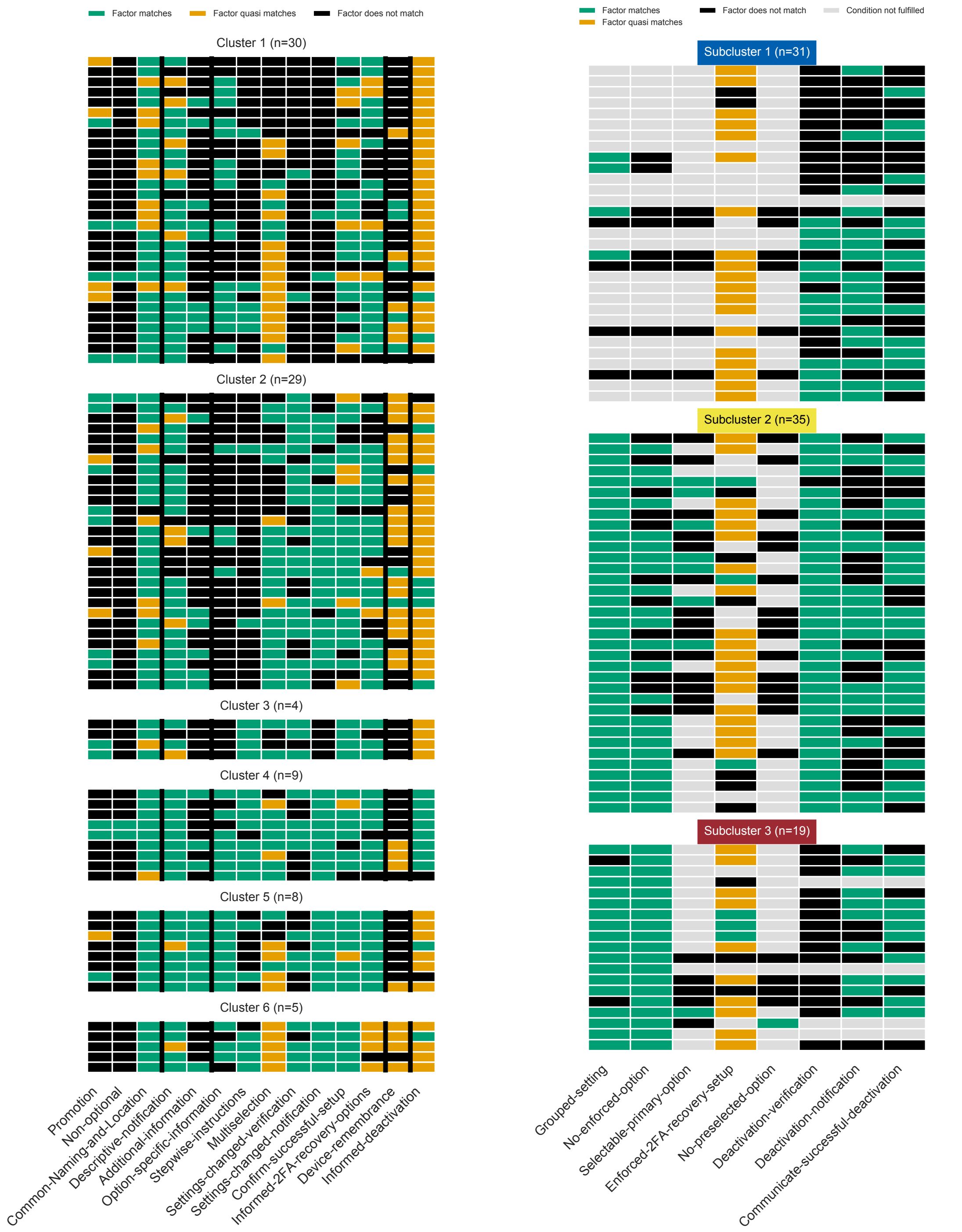


Figure 4.13: Clusters of websites based on comparison factors. Subclusters based on the conditional factors are indicated in the first column. Thick lines separate factors from different steps in the user journey (see also Table 4.1).



(a) Clusters of websites based on non-conditional factors. Only non-conditional factors are shown.

(b) Subclusters of websites based on conditional factors. Only conditional factors are shown.

Figure 4.14: Clusters for only non-conditional factors and for only conditional factors.

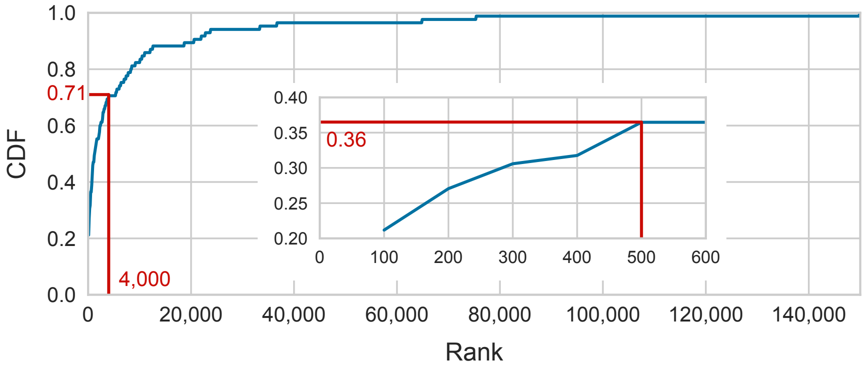


Figure 4.15: Cumulative frequency distribution of Tranco (125, 184) rankings of our 85 websites. Percentiles for 0.36 (31 websites) and 0.71 (60 websites) are marked, which correspond to websites in the top-500 and in the top-4000 in Tranco.

	Rank category			
Cluster	Top-500	Top-4000	Long tail	$\sum$
1	6	10	14	30
2	13	9	7	29
3	2	1	1	4
4	2	6	1	9
5	3	3	2	8
6	5	0	0	5
$\sum$	31	29	25	85

Table 4.5: Contingency table for cluster vs. rank category.

popular websites in each category, this distribution is naturally heavily skewed toward the top ranks. We then used the *inter*-cluster to describe each website's underlying 2FA user flow, which we analyzed for an association with the website Tranco rank group. Table 4.5 shows the contingency table for cluster vs. rank and Figure 4.16 illustrates the normalized contingency table. Fisher's exact test ($p = 0.04388$) shows that this association is statistically significant.

We also considered the association between website categories and clusters, but unfortunately, the website categories are too diverse, and the number of websites per category is too small to derive a meaningful connection between cluster and category. We are also unaware of any reliable, more coarse-grained website categorization that could be used.

4.7.2.4 Opinionated Separation of Comparison Factors

Our analysis considered all the comparison factors at once and did not differentiate between different categories of factors. To provide a different view on the consistency of 2FA user journeys, we conducted an expert evaluation of our factors to create an

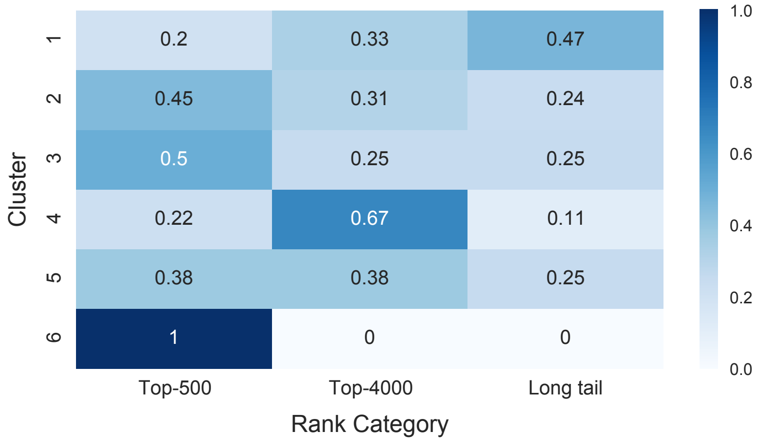


Figure 4.16: Heatmap of the normalized contingency table (see Table 4.5) for website cluster vs. website rank category.

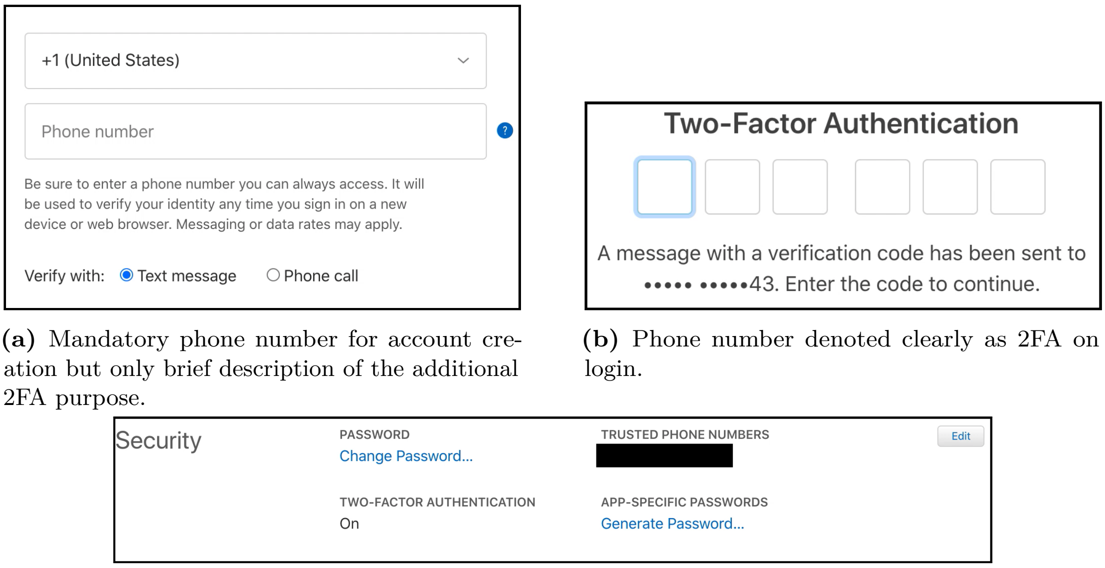
opinionated separation of factors by their impact on security, user experience, both, or neither. The entire evaluation process is described in Appendix C.5 and Table 4.6 summarizes the results. As a result, we split our comparison factors into four disjoint sets: *Non-conditional-UX* (7 factors), *Non-conditional-Security* (6 factors), *Conditional-UX* (5 factors), and *Conditional-Security* (3 factors). Only the factor *Additional-information* was considered irrelevant for UX or security. Based on the four sets of factors, we repeated the data analysis of Sections 4.7.2.1 and 4.7.2.2.

Pairwise Hamming distance: Considering only *non-conditional-UX* factors, the average website differs in 3–4 of the 7 factors from other websites, and considering only *non-conditional-security* factors, the average website differs in 2–3 of the 6 factors from other websites. Thus, with this distance metric, the websites in our data set do not show better consistency when considering separated sets of factors.

Factor clusters: For each set of factors, we calculate the mean Silhouette coefficient for different numbers of clusters with KModes to determine the best number of clusters to describe our data set. Compared to clustering with all factors, we found that the best-fitting number of clusters is larger when considering our separated factors. For *Non-conditional-UX* comparison factors, we found 5 clusters, and for *Conditional-UX* comparison factors, 10 clusters. For *Non-conditional-security* comparison factors, we calculated 9 clusters as the best number of clusters. For *Conditional-security* comparison factors, Silhouette testing showed 8 to be the best number of clusters. As a result, considering sets of separated factors, we found more diverse strategies for how websites in our data set implement their 2FA user journeys with regard to purely UX or security. Appendix C.5.2 illustrates the corresponding clusters.

Category	Conditional	Factors
User Experience	No	Promotion
	No	Common-Naming-and-Location
	No	Descriptive-Notification
	No	Option-Specific-Information
	Yes	Grouped-Setting
	Yes	Selectable-Primary-Option
	No	Confirm-Successful-Setup
	No	Informed-2FA-Recovery-Options
	Yes	Enforced-2FA-Recovery-Setup
	No	Device-Remembrance
	Yes	No-Preselected-Option
	Yes	Communicate-Successful-Deactivation
Security	No	Settings-Changed-Verification
	Yes	Deactivation-Verification
Both Security and User Experience	No	Non-optional
	No	Step-Wise-Instructions
	No	Multiselection
	Yes	No-Enforced-Options
	No	Settings-Changed-Notification
	No	Informed-Deactivation
	Yes	Deactivation-Notification
Neither	No	Additional-Information

Table 4.6: Separation of comparison factors.



+1 (United States)

Phone number

Be sure to enter a phone number you can always access. It will be used to verify your identity any time you sign in on a new device or web browser. Messaging or data rates may apply.

Verify with: ● Text message ○ Phone call

(a) Mandatory phone number for account creation but only brief description of the additional 2FA purpose.

Two-Factor Authentication

A message with a verification code has been sent to ••••• •••••43. Enter the code to continue.

(b) Phone number denoted clearly as 2FA on login.

Security

PASSWORD
Change Password...

TRUSTED PHONE NUMBERS
[redacted]

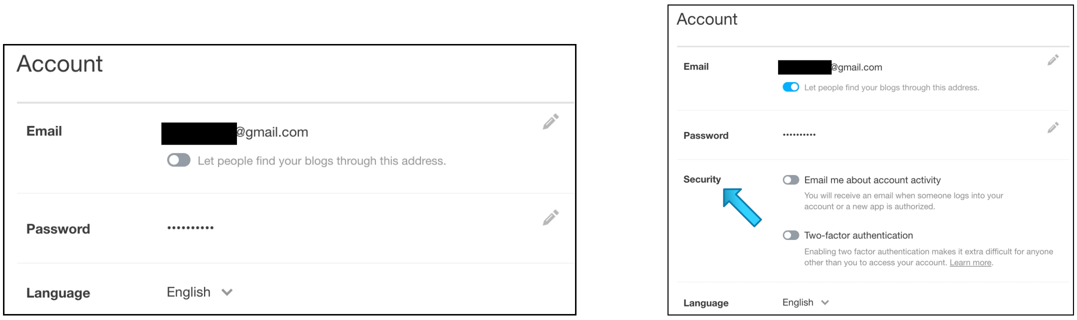
Edit

TWO-FACTOR AUTHENTICATION
On

APP-SPECIFIC PASSWORDS
Generate Password...

(c) Settings do not allow change or deactivation of 2FA.

Figure 4.17: Part of 2FA user journey of `icloud.com`.



Account

Email [redacted]@gmail.com
Let people find your blogs through this address.

Password ••••••••••

Language English

(a) `tumblr.com` before email verification

Account

Email [redacted]@gmail.com
Let people find your blogs through this address.

Password ••••••••••

Security
Email me about account activity
You will receive an email when someone logs into your account or a new app is authorized.
Two-factor authentication
Enabling two factor authentication makes it extra difficult for anyone other than you to access your account. Learn more.

Language English

(b) `tumblr.com` after email verification

Figure 4.18: `tumblr.com` has a *Common-Naming-and-Location* (●), but the security settings are initially hidden until the user verifies their email address.

4.7.3 Qualitative Data Analysis

We discuss the consistencies and inconsistencies we observed during our analysis of the 2FA user journeys.

4.7.3.1 Consistent Discovery for Self-Motivated Users

Our analysis shows that the vast majority of websites in our data set did not *immediately* promote 2FA to their end users in any form before/during sign-up and login—a website might promote 2FA only at a later point (e.g., an account existed for some time or the user takes actions that increase the severity of an account compromise), which our recording of journeys does not cover. The few websites that immediately promoted their 2FA support did this with mixed strategies, where most of them promoted 2FA during or immediately after account creation. In contrast, the remaining websites mentioned it only on their landing page, where users could easily miss it. However, we discovered that some websites' nudging to 2FA merely redirected the user to the account settings' security section, where the user has to pick up the journey themselves. Furthermore, six websites in our data set mandated 2FA, most of those sites in the cryptocurrency category. However, for two websites that mandate 2FA, we found that the intention to use the phone number or verified email address as a second factor was not clearly communicated to the user during account creation (e.g., Figure. 4.17).

Our analysis showed that users looking for 2FA settings have a consistent experience across websites. Almost all websites used a common location for their 2FA settings. Therefore, users who once went through the 2FA workflow can find the 2FA settings more easily on other websites. Most websites also use similar descriptions of 2FA (e.g., "second layer of security," "prevent unauthorized access," or "ask for authentication on new devices"), which helps the user to recognize the 2FA settings despite variations in the naming. Examples of clear exceptions to this pattern are illustrated in Figure 4.18 and Figure 4.3d.

4.7.3.2 Consistent Lack of Informing and Educating Users

We found that only a minority of the websites provided additional information (e.g., “learn more” link to detailed information including pictures and tutorials), and even fewer websites educate the user about the benefits and drawbacks of the 2FA options that they support, but instead immediately start the setup process. During this setup, only about a third of all websites guided the user with step-by-step instructions for setting up a chosen 2FA option. Most websites require the user to verify their identity to change their 2FA settings and inform them about such changes (e.g., by email). Very noticeable exceptions are the websites in Cluster 1, which almost entirely omit both verification and notification of settings changes.

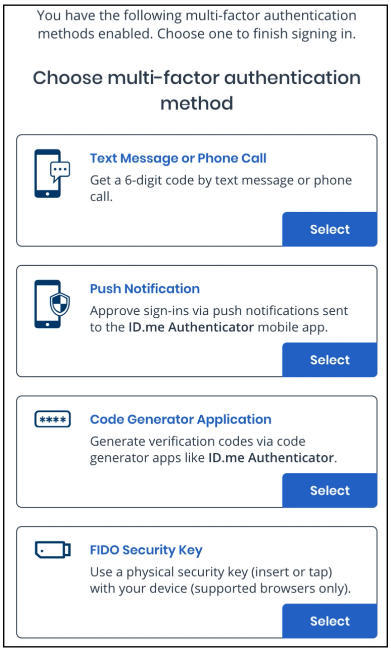
The most consistent behavior we have observed to inform users is the confirmation of a successful setup. More than four-fifths of the websites required a positive confirmation from the user (e.g., the user had to enter the current TOTP code to complete the setup) and, in most cases, also provided some visual feedback to the user to inform them about the successfully concluded 2FA setup.

4.7.3.3 Mixed Strategies for 2FA Setup and Configuration

We observed the most inconsistent behavior when it comes to the setup of 2FA options and their possible configurations by the user. First, there is an almost even split between three basic strategies: “offering only one 2FA option,” “offering multiple 2FA options but only one can be active at a time,” and “offering multiple 2FA options and multiple can be active at the same time.” Unfortunately, we could not find an explanation on any of the websites that let their users select only one 2FA option about why they implement 2FA this way. Second, among the websites that support multiple 2FA options, all but six websites show the 2FA options grouped in the same settings location, while those six exceptions, for instance, differentiate between 2FA and security keys in their security settings. However, thirdly, half of the websites with support for multiple 2FA options enforce a particular 2FA option to be set up before they offer the other options to the user. For example, only after providing their phone number can the user set up security keys or TOTP as an alternative. Fourth and last, websites are very consistent in proposing the 2FA option that should be used to login. Very few websites allow the user to select the 2FA option that should be used primarily for the login. Only a single website asked the user upfront during login which 2FA option they would like to use for the current login (see Figure 4.19). The vast majority of websites used internal metrics to determine which 2FA option should be used for login, and the user could only navigate through the “use a different method” or “do you have difficulties” menus to select another 2FA option.

4.7.3.4 Mainly Optional Recovery

Three-quarters of all websites offer recovery options, and most of those websites also explain to the user the importance of setting up recovery options or the risks of neglecting to set up a recovery option. The preferred recovery option among these websites was printable one-time codes. Also, websites are very consistent in enforcing the setup



You have the following multi-factor authentication methods enabled. Choose one to finish signing in.

Choose multi-factor authentication method

Text Message or Phone Call

Get a 6-digit code by text message or phone call.

Push Notification

Approve sign-ins via push notifications sent to the **ID.me Authenticator** mobile app.

Code Generator Application

Generate verification codes via code generator apps like **ID.me Authenticator**.

FIDO Security Key

Use a physical security key (insert or tap) with your device (supported browsers only).

Figure 4.19: `id.me` allows users to choose their 2FA option upfront during login (*No-preselected-option*: ●) and supports multiple activated 2FA options (*Multiselection*: ●).

of a recovery option. Almost three-quarters of the websites with a recovery option nudge the user to set up the recovery, and only six websites enforce this during the 2FA setup. This low number of websites with mandatory recovery also means that there is no fail-safe account recovery strategy by websites that support at most one active 2FA option. It would be intuitive that such websites would enforce a recovery option to prevent account lockout in case this single option is unavailable. Still, our data set does not support this. For the usage of recovery options, we found only one website that, although supporting one-time codes, does not offer an obvious way to use them (i.e., there was no link to a recovery page and no instructions to use the recovery codes as input to the regular OTP form field).

4.7.3.5 Mixed Strategies for Device Remembrance

More than half of the websites in our data set do not support device remembrance, i.e., the user cannot explicitly select to skip the second-factor authentication during future logins. For the websites that support this feature, we found that not only do they describe it in different ways but also that their remembrance logic differs. Almost two-thirds of the websites need the user to opt-*in* to this feature, a fifth of the websites needs the user to explicitly opt-*out* from remembering the device, and another fifth of the websites automatically places a device remembrance cookie without asking the user.

4.7.3.6 Consistent Support for Deactivation

All but five websites in our data set support the deactivation of 2FA (and only two of those exceptions mandate the setup of 2FA). However, only a minority of websites communicate to the user the risk of deactivating 2FA (e.g., easier account hijacking). Furthermore, similar to the previously mentioned lack of consistently informing and educating users about 2FA, we find that only about half of all websites verify the user identity before deactivation, notify the user about deactivation, or communicate a successful deactivation in the website settings.

4.8 Discussion and Future Work

Are the websites in our data set consistently following the same design patterns and strategies in offering 2FA? Although some factors individually show high consistency, we did not find a single start-to-end design pattern for the user journey that is consistently followed by the majority of websites in our data set. Instead, we found that 2FA user journeys are clustered into smaller groups of websites with similar journeys. Separating the comparison factors by their impact on UX and security did not indicate a more consistent strategy for pure usability or security-related steps along the user journey. In fact, we found that the separated sets of factors were more clustered. Taking into account the rankings of websites, our results indicate that the design represented by clusters 2, 3, and 6 is more popular among the top-ranked websites, while more than 50% of the websites from the long-tail ranks are in cluster 1.

Implications for developers and users: UX guidelines state that users prefer a site to work in the same way as all other sites they already know. This heuristic has been shown to be successful on the Web, for example, when it comes to online shopping or banking experiences. To follow this heuristic vis-a-vis two-factor authentication as a factor for the overall user experience on the Web, developers could follow the 2FA experience on the majority of websites or on the most popular websites, which likely minted the users' mental models. Our results show that such a majority does not exist among the popular websites and that even the leading websites (e.g., Google and Apple) do not agree in their user journeys. Therefore, a recommendation for influential industry associations and consortia would be to draft recommendations for website developers on how to achieve a consistent strategy for 2FA user journeys. Possible avenues for the community and future work to contribute to this endeavor could be to create guidelines that foster consistent strategies for implementing the best possible 2FA UX on the Web.

A crucial consideration when striving for consistency is that consistency in itself does not guarantee a good UX; a bad design could be consistently implemented, but users have learned to live with it. As a concrete example from our data set, Apple's `icloud.com` is an outlier in various comparison factors: it mandates a phone number as the only 2FA option without clearly informing the user during account creation and without the option to add other options later or to deactivate it (see Figure 4.17). But do users conceive `icloud.com`'s 2FA user journey as a bad or good experience? Our study design does not attempt to assign a quality measurement to individual factors, and it does not measure the quality of user experiences attached to the different clusters of user journeys. But clearly, our results motivate that the impact of the different strategies for the 2FA user journey on the perceived usability by users has to be thoroughly investigated in an effort to make the best strategy consistent across websites. Our data indicated a connection between the rank of a website and the site's strategy, but it is unclear to what extent the guidelines for 2FA journeys need to be contextualized. For example, certain comparison factors may be dictated or recommended by regulations, such as PCI-DSS or the EU Revised Payment Services Directive (PSD2) with strong customer authentication (SCA), different types of websites may have different security policies, or specific user groups [39, 140] require different support. Thus, it is unclear whether consistency between *all* comparison factors is required or even desirable.

Indications for the external validity of prior works: Although we did not study the usability of individual instruments, comparison factors, or steps in the user journey, we can provide a new perspective on some aspects of previous work and UX guidelines we observed during our data analysis. We noted that the discovery of 2FA and the initial education of users are very consistent and that there is a common naming and description of 2FA in place. But neither of the two types of 2FA descriptions that we have noted in our analysis (see Table 4.2) complies with recent results by Golla et al. [82] and Lassak et al. [120] on how users should be nudged and educated to encourage the adoption of 2FA. Furthermore, Ciolino et al. [32] conducted a user study of 2FA setup and login "in the wild." Their participants encountered some of the patterns we identified in our work and described them as problematic. For instance, enforcing the SMS 2FA option while not communicating that additional 2FA options become available after registering the phone number confused participants that were explicitly looking for registration of security keys; an opt-out device remembrance, which we found on several websites (7×●, 8×◐), frustrated participants that were expecting to be prompted for a second factor on login but missed that they had to take explicit action for that; and their participants expressed the desire for personalization by being able to select the preferred 2FA option for logins, which we found is not a widespread feature but, on the contrary, the 2FA option is in most cases chosen by the website (only 8 out of 34 websites with fully matched *Multiselection* allowed setting a primary option, and only one website of the remaining 26 sites did *not* pre-select the option). Lastly, from our clustering, we noticed that recommendations by UX guidelines to provide adequate contextual help and break down complex tasks, in this case for setting up 2FA, were ignored on many websites that did not offer additional or option-specific information or simply step-wise instructions. Also, the recommendation to provide clear feedback to users was not realized on many websites that did not notify users or communicate a successful 2FA setup or deactivation. Thus, our study provides indications for the external validity of prior results. In our opinion, measuring to what extent each pattern we detected matches the recommendations and settings of related work would be an interesting follow-up study to provide better insights into the external validity of previous studies (e.g., taking textual content and UI designs into more consideration). Those indications also emphasize the need to establish more general UX guidelines for implementers of 2FA user journeys to improve the usability of 2FA. The first option-specific guidelines [72, 73] or collections of best-practices [51, 198] are a good starting point.

FIDO UX Guidelines [72, 73]: The FIDO Alliance UX guidelines also consider similar steps in the user journey (promotion, invitation, registration, and login). They recommend the promotion of biometric awareness or security keys at sign-in and registration, educating users about the FIDO value proposition of a "simple and secure sign-in without password" or about authentication with security keys, providing a "learn more" link and giving concrete statements based on user studies, confirming successful registration with a clear indication to users, encouraging users to register mulitple keys for recovery and backup [73], and explicitly promoting "Security and Privacy settings" to manage 2FA options. Unfortunately, these guidelines are not suitable as a general guideline and, in some points, conflict with recent recommendations from research (e.g., the promotion message [120, 82] or automatically setting FIDO2 as the default sign-in option [32]). The guidelines [72] are strongly tailored to promote biometric authentication as a convenient alternative to passwords or to promote 2FA with security keys to consumers of regulated industry websites [73], such as banking or healthcare. They do not target a 2FA setting [72], and the guidelines do not address the setup and UX of multiple authentication options, or they limit themselves to only security keys as the second factor [73]. For desktop authenticators [72], the password is considered a fallback option; therefore, these guidelines omit explicit recovery steps.

Limitations: Like any other qualitative study, our work also has some limitations. Despite our best efforts, we cannot exclude a subjective bias by the involved researchers, e.g., in identifying the comparison factors or selecting a clustering with the best descriptive power. We aimed to study 120 popular websites, but only 85 were possible due to various restrictions and obstacles. Thus, our study is skewed toward top websites in English and from specific categories. We fixed the conditions for data collection to increase the internal validity of our data, but we cannot exclude that our setting is considered high-risk or low-risk by a website and that we experienced a different user journey than other users of the same site. Moreover, we collected our data only from desktop computers; thus, our comparison factors may differ on mobile devices. Furthermore, with the adoption of new technologies (e.g., WebAuthn) and changes in website policies (e.g., Google plans to mandate 2FA for an increasing number of its users [130]), our comparison might not capture the most recent picture. However, we believe that our general results remain valid. Lastly, we did not continue to monitor the websites, nor did we explore the user flow for going through account recovery or changing 2FA-relevant information (e.g., phone number or email address), since we focused on the steps of the user journey that mint the users' initial impressions of 2FA.

Future work: Conducting user and developer studies is an obvious path to follow up on our results. While we detected a lack of consistency in the 2FA user journeys that can increase users' cognitive friction, it is unclear if this contributes to the notoriously low adoption rate of 2FA among end-users. Our survey indicated that several users did indeed refrain from setting up 2FA or deactivated it due to differences in the user experience between websites. Furthermore, it is unclear whether a "gold standard" for journeys exists or to what extent journeys need to be contextualized (e.g., website category, regulations, or specific user groups). Comparative studies of different design patterns could answer those questions and others, like a weighting of comparison factors by their impact on, e.g., the UX or 2FA security. Moreover, we consider it worthwhile to explore developers' reasons for choosing a particular design pattern to understand the reasons behind those inconsistent journeys. In addition to human-centered studies, extending our methodology to user journeys for account recovery, to other device form factors, such as mobile devices, or to entirely new solutions, such as Passkey, would complement our results. Lastly, we think that studying the 2FA user journeys can provide insights into the external validity of (prior) studies of individual aspects of 2FA and shed new light on what constitutes a good 2FA user experience.

4.9 Related Work

Several works have studied two-factor authentication problems and focused on the usability component and user attitudes. Bonneau et al. [15] conducted a systematic expert assessment of various authentication solutions, including many of the solutions used for 2FA. They concluded that the usability of these solutions falls very often short compared to text-based passwords. In contrast to Bonneau et al., most other works relied on user studies to investigate problems of 2FA.

A focal point of prior user studies was the setup and usage of different two-factor authentication solutions to understand users' attitudes toward 2FA, obstacles for its adoption, and how to improve the usability and user experience. Previous works studied two-factor authentication in settings such as online banking [201, 202, 86, 114] or military [180] services. Like other studies of 2FA [44, 19, 63, 45] they found that users consider 2FA often burdensome and slow, that convenience trumps perceived security, and that users do not always understand the risks that 2FA tries to remedy. Several works have studied 2FA problems in organizational contexts [182, 4, 33, 163, 56, 197] where the use of MFA can be mandated. While those studies show that many of the problems overlap with non-organizational settings, they could also shed new light on the positive influence of features such as device remembrance [163, 56] or better help and instructions.

Several studies [161, 201, 202, 114] compared different options for the second factor to identify option-specific differences in user attitudes and usability, while other works specifically studied security keys [38, 164, 32] or authenticator apps [43]. An interesting aspect of those works [161, 32, 164] for our study is that they differentiated between 2FA setup and login, where users often struggled in the setup due to unclear instructions/workflows. Strong recommendations from those works were clearer instructions and guidance for the setup to avoid user frustration that often leads to non-adoption.

Additionally, improved notification design patterns [82] have been shown to encourage users to adopt 2FA.

Lastly, recent works [P2, 65, 151, 120] studied specifically FIDO2 *single*-factor authentication. They found similar user concerns as for 2FA. However, the weighting of the concerns shifted (e.g., loss of the authenticator device is ranked very high) or new concerns were added (e.g., misunderstanding biometric WebAuthn). Relevant to our work, the FIDO Alliance has recently published UX guidelines for security keys [73] and implementers of desktop authenticators [72] that, similar to our methodology, divide the user journey into different steps and provide recommendations for the design of each step; however, explicitly tailored to the technical details of FIDO2/WebAuthn with biometric authenticator devices or security keys. Nevertheless, those guidelines incorporate many of the UX guidelines explained in Section 4.3.2.

The key difference of our work is that we do not study how concrete changes in form, content, or functionality affect the usability and concrete experience of 2FA, but that we are first to systematically study how *consistent* the user experience is across existing popular websites. Our work, in contrast to previous works, strongly focuses on Jakob's law of Internet user experience which states that an inconsistent user experience across websites increases cognitive friction and can be detrimental to users' adoption. Providing first insights into how well the 2FA user journeys adhere to this law is the core contribution of this work. Further, we are not aware of prior studies that measured Jakob's law across a larger number of websites but instead, to the best of our knowledge, qualitative and quantitative testing of websites focuses on single websites or comparative user studies between a small set of websites based on general UX best-practices and guidelines. Therefore, we had to devise a methodology to measure the consistency of the 2FA user journeys on different websites.

4.10 Conclusion

This work contributes a methodology for comparing 2FA user journeys on websites and presents the first systematic study of the consistency of those journeys on top-ranked websites. Our results show a lack of consistency for the various steps along those journeys. We find that even the more consistent design patterns were described as problematic for usability in the literature. We strongly believe that our results motivate different future works that can lead to the creation of more general user experience guidelines for implementers of two-factor authentication.

5
Conclusion

Authentication is one of the most fundamental security goals for protecting online services. Only with reliable authentication can it be ensured that online services and their related data can solely be accessed by authorized persons. For decades, text-based passwords have been the incumbent user authentication method on the web. However, research and practice have shown over and over again that regular users are not capable of creating truly unique, hard-to-guess passwords. This has several reasons, such as the increasing digitalization of users' daily lives leading to an ever-increasing number of user accounts that require authentication. This leads to password reuse exploited in credential stuffing attacks and easy-to-crack passwords despite secure password storage. To assist users in creating "strong" passwords, they are usually referred to password manager programs as a technical aid to create, store, and use randomly generated passwords that are unique and virtually impossible to guess. But the security benefits of those password manager programs depend on the users consistently employing them to *generate* passwords. Only storing and conveniently entering the existing weak passwords of users does not improve the users' security hygiene. This thesis was the first to show that password managers do not have the desired effect in practice [P1]: users still store reused, easy-to-guess passwords in their password managers and rarely use those tools consistently for password generation. Our results, taken together with prior work [105, 29, 127, 220] that identified usability and trust issues of users with password managers, show that we need to turn to other solutions to improve the users' online authentication hygiene effectively.

Currently, web authentication appears to be at a turning point: the number of websites supporting two-factor authentication increases steadily and the industry, with the FIDO Alliance leading the way, pushes for token-based authentication, even in a passwordless setting. However, this turning point also means a paradigm shift for users. After being lectured and urged for decades about their password security, all of a sudden, users must develop an understanding and the right mental model for two-factor authentication, and token-based authentication in particular. In this thesis, we were the first to study this paradigm shift [P2] for passwordless authentication with security keys. We found that while users consider FIDO2 passwordless authentication as more usable and more acceptable than traditional passwords, concerns remain that impede many users' willingness to abandon passwords. This is rooted in a gap between the user's *personal* perspective on this new technology and the global view of the FIDO2 designers that might not sufficiently include the users' views. Different related works (e.g., [65, 151, 120, 117, 107]) that followed our study re-confirmed our results and/or identified additional hurdles for user adoption. Since the publication of our results, the FIDO alliance has released its passkey specification, a way to synchronize the users' private keys (usually stored on a token device) across their personal devices using, for example, the Apple, Google, or Microsoft ecosystems. Passkeys currently receive a lot of attention from the industry, and there is an observable push to adopt this technology in online services. While this new concept addresses some of the concerns raised in our and others' studies, we think that the road to creating the right mental models for users and fully adopting FIDO2/Passkeys still appears to be long and requires more research about how to help users to adjust to this new paradigm.

In contrast to Passkeys and passwordless WebAuthn, two-factor authentication has been supported by websites for a longer time. Two-factor authentication can come in different forms, such as SMS, TOTP from apps or tokens, push notifications, or, most recently also, FIDO2/Webauthn. Yet, despite being supported for a longer time, adoption by users has been notoriously low. In this thesis, we contributed a new perspective on the adoption of 2FA in general [P3]. Consistency of user journeys is an established UX heuristic that aims at decreasing the cognitive friction that users experience when using a new product by allowing users to transfer their prior experience. We were the first to investigate the consistency of the UX experience across different websites using a novel methodology. Those results show a lack of consistency for the various steps along those journeys and that even the more consistent design patterns were described as problematic for usability in the literature. Our results motivate different future works that can lead to the creation of more general user experience guidelines for implementers of two-factor authentication, including user studies of different design patterns, developer studies about their reasons for choosing a particular design pattern, or extending our methodology to account recovery, other device form factors, or new solutions, such as Passkey.

Overall, our results underline that technical advances for solving strong user authentication on the web need to be better accompanied by usability advances that are rooted in principled research rather than assumptions about users (e.g., that users by default use password managers correctly or that users naturally accept biometric authentication to web services and security keys because they are easier and more convenient than passwords). Without such accompanying advances in usability research, we predict that even highly promising solutions, like Passkeys, may be doomed to fail in practice because of lacking user acceptance. Developing new authentication solutions for users must become a human-centered discipline that first considers the user at all development stages.

Bibliography

Author's Papers for this Thesis

[P1] Ghorbani Lyastani, S., Schilling, M., Fahl, S., Bugiel, S., and Backes, M. Studying the impact of managers on password strength and reuse. In: *Proc. 26th USENIX Security Symposium (SEC' 18)*. USENIX Association, 2018.

[P2] Ghorbani Lyastani, S., Schilling, M., Neumayr, M., Backes, M., and Bugiel, S. Is FIDO2 the kingslayer of user authentication? A comparative usability study of FIDO2 passwordless authentication. In: *Proc. 41st IEEE Symposium on Security and Privacy (SP '20)*. IEEE, 2020.

[P3] Ghorbani Lyastani, S., Backes, M., and Bugiel, S. A systematic study of the consistency of two-factor authentication user journeys on top-ranked websites. In: *Proc. 30th Annual Network and Distributed System Security Symposium (NDSS '23)*. The Internet Society, 2023.

Other Papers of the Author

[S1] Ghorbani Lyastani, S., Acar, Y., Backes, M., and Fahl, S. Poster: Improving password memorability and strength using mangling rules. In: *Proc. 12th Symposium on Usable Privacy and Security (SOUPS'16)*. USENIX Association, 2016.

Other references

[1] *1Password*. `https://1password.com/`.

[2] *2FA Directory*. `https://2fa.directory/`. Feb. 25, 2021.

[3] *2factorauth*. `https://github.com/2factorauth/twofactorauth`. Feb. 25, 2021.

[4] Abbott, J. and Patil, S. How mandatory second factor affects the authentication user experience. In: *Proc. CHI Conference on Human Factors in Computing Systems (CHI'20)*. ACM, 2020.

[5] Ackermann, Y. *WebAuthn Awesome: A curated list of awesome WebAuthn/FIDO2 resources*. `https://github.com/herrjemand/awesome-webauthn`. 2019.

[6] Aebischera, S., Dettoni, C., Jenkinson, G., Krol, K., Llewellyn-Jones, D., Masui, T., and Stajano, F. Pico in the wild: replacing passwords, one site at a time. In: *2nd European Workshop on Usable Security (EuroUSEC '17)*. 2017.

[7] *Alexa Web Information Service: Developer Guide (API Version 2005-07-11)*. `https://docs.aws.amazon.com/AlexaWebInfoService/latest/`.

[8] All about UX. *User experience definitions*. `http://www.allaboutux.org/ux-definitions`.

[9] Apple Inc. *Human Interface Guidelines*. `https://developer.apple.com/design/human-interface-guidelines/`.

[10] Attig, C., Wessel, D., and Franke, T. Assessing personality differences in human-technology interaction: an overview of key self-report scales to predict successful interaction. In: *HCI International 2017 – Posters' Extended Abstracts*. 2017.

[11] Bailey, D. V., Dürmuth, M., and Paar, C. Statistics on password re-use and adaptive strength for financial accounts. In: *Proc. 9th International Conference on Security and Cryptography for Networks (SCN'14)*. 2014.

[12] Bangor, A., Kortum, P., and Miller, J. Determining what individual sus scores mean: adding an adjective rating scale. *J. Usability Studies* 4, 3 (May 2009).

[13] Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., and Green, P. *lme4: Linear Mixed-Effects Models using 'Eigen' and S4*. `https://cran.r-project.org/web/packages/lme4/index.html`.

[14] Bonneau, J. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In: *Proc. 33rd IEEE Symposium on Security and Privacy (SP '12)*. IEEE, 2012.

[15] Bonneau, J., Herley, C., Oorschot, P. C. v., and Stajano, F. The quest to replace passwords: a framework for comparative evaluation of web authentication schemes. In: *Proc. 33rd IEEE Symposium on Security and Privacy (SP '12)*. IEEE, 2012.

[16] Bonneau, J. and Preibusch, S. The password thicket: technical and market failures in human authentication on the web. In: *9th Workshop on the Economocs of Info Security (WEIS'10)*. 2010.

[17] Brand, C. *Advisory: Security Issue with Bluetooth Low Energy (BLE) Titan Security Keys*. `https://security.googleblog.com/2019/05/titan-keys-update.html`. May 2019.

[18] Brand, C. and Kitamura, E. *Enabling Strong Authentication with WebAuthn*. `https://developers.google.com/web/updates/2018/05/webauthn`. 2019.

[19] Braz, C. and Robert, J.-M. Security and usability: the case of the user authentication methods. In: *Proc. 18th Conference on L'Interaction Homme-Machine (IHM '06)*. ACM, 2006.

[20] Breusch, T. S. and Pagan, A. R. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society* (1979), 1287–1294.

[21] Brooke, J. SUS—a quick and dirty usability scale. *Usability Evaluation Industry* 189, 194 (Nov. 1996), 4–7.

[22] Brunswick, S. Ecommerce fraud – time to act. *Card Technology Today* 21, 1 (2009), 12–13.

[23] Bursztein, E. *The bleak picture of two-factor authentication adoption in the wild.* https://elie.net/blog/security/the-bleak-picture-of-two-factor-authentication-adoption-in-the-wild/. Dec. 2018.

[24] Carné de Carnavalet, X. de and Mannan, M. From very weak to very strong: analyzing password-strength meters. In: *Proc. 21th Annual Network and Distributed System Security Symposium (NDSS '14).* The Internet Society, 2014.

[25] Castelluccia, C., Dürmuth, M., and Perito, D. Adaptive password-strength meters from markov models. In: *Proc. 19th Annual Network and Distributed System Security Symposium (NDSS '12).* The Internet Society, 2012.

[26] Ceti, S. *The password is dead, long live Web Authentication.* https://www.computerworld.com.au/article/647205/password-dead-long-live-web-authentication/. Sept. 2018.

[27] Chatterjee, R., Doerfler, P., Orgad, H., Havron, S., Palmer, J., Freed, D., Levy, K., Dell, N., McCoy, D., and Ristenpart, T. The spyware used in intimate partner violence. In: *Proc. 39th IEEE Symposium on Security and Privacy (SP '18).* IEEE, 2018.

[28] Chaturvedi, A., Green, P. E., and Caroll, J. D. K-modes clustering. *Journal of classification* 18, 1 (2001), 35–55.

[29] Chiasson, S., Oorschot, P. C. van, and Biddle, R. A usability study and critique of two password managers. In: *Proc. 15th USENIX Security Symposium (SEC '06).* USENIX Association, 2006.

[30] Chong, J. *10 Things You've Been Wondering About FIDO2, WebAuthn, and a Passwordless World.* https://www.yubico.com/2018/08/10-things-youve-been-wondering-about-fido2-webauthn-and-a-passwordless-world/. Aug. 2018.

[31] Christensen, R. H. B. *ordinal: Regression Models for Ordinal Data.* https://cran.r-project.org/web/packages/ordinal/index.html.

[32] Ciolino, S., Parkin, S., and Dunphy, P. Of two minds about two-factor: understanding everyday FIDO u2f usability through device comparison and experience sampling. In: *Proc. 15th Symposium on Usable Privacy and Security (SOUPS'19).* USENIX Association, 2019.

[33] Colnago, J., Devlin, S., Oates, M., Swoopes, C., Bauer, L., Cranor, L., and Christin, N. "it's not actually that horrible": exploring adoption of two-factor authentication at a university. In: *Proc. CHI Conference on Human Factors in Computing Systems (CHI'18).* ACM, 2018.

[34] Conners, J. S. and Zappala, D. Let's Authenticate: Automated Cryptographic Authentication for the Web with Simple Account Recovery. In: *Who Are You?! Adventures in Authentication Workshop (WAY '19)*. 2019.

[35] Cooper, A., Reimann, R., Cronin, D., and Noessel, C. *About Face: The Essentials of Interaction Design*. 4th. Wiley John + Sons, 2014.

[36] D. Van Der Laan, J., Heino, A., and Waard, D. A simple procedure for the assessment of acceptance of advance transport telematics. *Transportation Research Part C: Emerging Technologies* 5 (Feb. 1997), 1–10.

[37] Das, A., Bonneau, J., Caesar, M., Borisov, N., and Wang, X. The tangled web of password reuse. In: *Proc. 21th Annual Network and Distributed System Security Symposium (NDSS '14)*. The Internet Society, 2014.

[38] Das, S., Dingman, A., and Camp, L. J. Why johnny doesn't use two factor: a two-phase usability study of the fido u2f security key. In: *Proc. Financial Cryptography and Data Security (FC'18)*. 2018.

[39] Das, S., Kim, A., Jelen, B., Huber, L., and Camp, L. J. Non-inclusive online security: older adults' experience with two-factor authentication. In: *Proc. 54th Hawaii International Conference on System Sciences*. 2020.

[40] Das, S., Kim, A., Jelen, B., Streiff, J., Camp, L. J., and Huber, L. Towards Implementing Inclusive Authentication Technologies for Older Adults. In: *Who Are You?! Adventures in Authentication Workshop (WAY '19)*. 2019.

[41] Das, S., Mare, S., and Camp, L. J. Smart storytelling: video and text risk communication to increase mfa acceptability. In: *Proc. 6th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 2020.

[42] Das, S., Russo, G., Dingman, A. C., Dev, J., Kenny, O., and Camp, L. J. A qualitative study on usability and acceptability of yubico security key. In: *Proc. 7th Workshop on Socio-Technical Aspects in Security and Trust (STAST '17)*. ACM, 2018.

[43] Das, S., Wang, B., and Camp, L. J. Mfa is a waste of time! understanding negative connotation towards mfa applications via user generated content. In: *Proc. 13th International Symposium on Human Aspects of Information Security & Assurance (HAISA'19)*. 2019.

[44] Das, S., Wang, B., Kim, A., and Camp, L. J. MFA is A necessary chore!: exploring user mental models of multi-factor authentication technologies. In: *Proc. 53rd Hawaii International Conference on System Sciences (HICSS'20)*. 2020.

[45] Das, S., Wang, B., Tingle, Z., and Camp, L. J. Evaluating user perception of multi-factor authentication: a systematic review. In: *Proc. 13th International Symposium on Human Aspects of Information Security & Assurance (HAISA'19)*. Springer, 2019.

[46] Dauterman, E., Corrigan-Gibbs, H., Mazières, D., Boneh, D., and Rizzo, D. True2f: backdoor-resistant authentication tokens. In: *Proc. 40th IEEE Symposium on Security and Privacy (SP '19)*. IEEE, 2019.

[47] Davis, F. D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 3 (Sept. 1989), 319–340.

[48] De Cristofaro, E., Du, H., Freudiger, J., and Norcie, G. A comparative usability study of two-factor authentication. In: *Workshop on Usable Security and Privacy (USEC'14).* The Internet Society, 2014.

[49] Dell'Amico, M., Michiardi, P., and Roudier, Y. Password strength: an empirical analysis. In: *Proc. 29th Conference on Information Communications (INFOCOM'10).* IEEE Press, 2010.

[50] Deveria, A. *Can I use WebAuthn?* May 2019. URL: `https://caniuse.com/#search=webauthn`.

[51] Dias, D. *9 Best Practices and UX Improvements for the two-factor authentication (2FA).* `https://thedaviddias.dev/blog/9-best-practices-ux-for-two-factor-authentification/`. Feb. 25, 2021.

[52] Doerfler, P., Thomas, K., Marincenko, M., Ranieri, J., Jiang, Y., Moscicki, A., and McCoy, D. Evaluating login challenges as adefense against account takeover. In: *The World Wide Web Conference (WWW '19).* ACM, 2019.

[53] Dragoljub, N. Stronger security. *Card Technology Today* 19, 1 (2007), 9–10.

[54] Dropbox. *Github: dropbox/zxcvbn.* `https://github.com/dropbox/zxcvbn`.

[55] Dürmuth, M., Angelstorf, F., Castelluccia, C., Perito, D., and Chaabane, A. Omen: faster password guessing using an ordered markov enumerator. In: *Proc. 7th International Symposium on Engineering Secure Software and Systems (ESSoS'15).* Springer, 2015.

[56] Dutson, J., Allen, D., Eggett, D., and Seamons, K. Don't punish all of us: measuring user attitudes about two-factor authentication. In: *Proc. European Symposium on Security and Privacy Workshops (EuroS&PW).* IEEE, 2019.

[57] Dynamo Wiki. *Guidelines for Academic Requesters (version 2.0).* `http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters`. Last visited: 11/10/17.

[58] E. Shannon, C. Prediction and entropy of printed english (1951).

[59] Egelman, S., Jain, S., Portnoff, R. S., Liao, K., Consolvo, S., and Wagner, D. Are you ready to lock? In: *Proc. 21th ACM Conference on Computer and Communication Security (CCS '14).* ACM, 2014.

[60] Egelman, S. and Peer, E. Scaling the security wall: developing a security behavior intentions scale (SeBIS). In: *Proc. CHI Conference on Human Factors in Computing Systems (CHI'15).* ACM, 2015.

[61] Egelman, S., Sotirakopoulos, A., Muslukhov, I., Beznosov, K., and Herley, C. Does my password go up to eleven?: the impact of password meters on password selection. In: *Proc. CHI Conference on Human Factors in Computing Systems (CHI'13).* ACM, 2013.

[62] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S. A geometric framework for unsupervised anomaly detection. In: *Applications of Data Mining in Computer Security*. Springer US, Boston, MA, 2002, 77–101.

[63] Fagan, M. and Khan, M. M. H. Why do they do what they do?: a study of what motivates users to (not) follow computer security advice. In: *Proc. 12th Symposium on Usable Privacy and Security (SOUPS'16)*. USENIX Association, 2016.

[64] Fahl, S., Harbach, M., Acar, Y., and Smith, M. On the ecological validity of a password study. In: *Proc. 9th Symposium on Usable Privacy and Security (SOUPS'13)*. ACM, 2013.

[65] Farke, F. M., Lorenz, L., Schnitzler, T., Markert, P., and Dürmuth, M. "you still use the password after all" – exploring fido2 security keys in a small company. In: *Proc. 16th Symposium on Usable Privacy and Security (SOUPS'20)*. USENIX Association, 2020.

[66] Farrell, S. *Open-Ended vs. Closed-Ended Questions in User Research*. https://www.nngroup.com/articles/open-ended-questions/. May 2016.

[67] *Fernando J. Corbató: Biography*. https://www.computer.org/profiles/fernando-corbato.

[68] FIDO Alliance. *Android Now FIDO2 Certified, Accelerating Global Migration Beyond Passwords*. Feb. 2019. URL: https://fidoalliance.org/android-now-fido2-certified-accelerating-global-migration-beyond-passwords/.

[69] FIDO Alliance. *Client to Authenticator Protocol (CTAP) — Proposed Standard, January 30, 2019*. Jan. 2019. URL: https://fidoalliance.org/specs/fido-v2.0-id-20180227/fido-client-to-authenticator-protocol-v2.0-id-20180227.html.

[70] FIDO Alliance. *FIDO Members*. 2019. URL: https://fidoalliance.org/members/.

[71] FIDO Alliance. *Recommended Account Recovery Practices for FIDO2 Relying Parties*. Feb. 2019.

[72] FIDO Alliance. *FIDO Desktop Authenticator UX Guidelines (v1)*. June 2021.

[73] FIDO Alliance. *FIDO Security Key UX Guidelines*. June 2022.

[74] Field, A. and Miles, J. *Discovering Statistics Using R*. Sage Publications Ltd., May 2012.

[75] Florencio, D. and Herley, C. A large-scale study of web password habits. In: *Proc. 16th International Conference on World Wide Web (WWW'07)*. ACM, 2007.

[76] Franke, T., Attig, C., and Wessel, D. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467.

[77] Frazier, S. *The 2019 State of the Auth Report: Has 2FA Hit Mainstream Yet.* https://duo.com/blog/the-2019-state-of-the-auth-report-has-2fa-hit-mainstream-yet. Dec. 2019.

[78] Gaw, S. and Felten, E. W. Password management strategies for online accounts. In: *Proc. 2nd Symposium on Usable Privacy and Security (SOUPS'06).* ACM, 2006.

[79] Ghorbani Lyastani, S. *Introduction Videos for FIDO2 1FA User Study.* Jan. 2020. URL: https://www.youtube.com/watch?v=_UCLBLo-fQI&list=PLJ3WH6JNU1HSyZKAfWtRLG0ONd5nsuABJ.

[80] Gioia, D. A., Corley, K. G., and Hamilton, A. L. Seeking qualitative rigor in inductive research: notes on the gioia methodology. *Organizational Research Methods* 16, 1 (2013), 15–31.

[81] Girardeau, B. *Introducing WebAuthn support for secure Dropbox sign in.* May 2018. URL: https://blogs.dropbox.com/tech/2018/05/introducing-webauthn-support-for-secure-dropbox-sign-in/.

[82] Golla, M., Ho, G., Lohmus, M., Pulluri, M., and Redmiles, E. M. Driving 2fa adoption at scale: optimizing two-factor authentication notification design patterns. In: *Proc. 30th USENIX Security Symposium (SEC' 21).* USENIX Association, 2021.

[83] Goodall, D. W. A new similarity index based on probability. *Biometrics* 22 (1966), 882.

[84] Google. *Material Design.* https://material.io/design.

[85] Grassi, P. A., Fenton, J. L., Newton, E. M., Perlner, R. A., Regenscheid, A. R., Burr, W. E., and Richer, J. P. *NIST SP800–63B: Digital authentication guideline (Authentication and Lifecycle Management).* Last visited: 30/05/2019. June 2017.

[86] Gunson, N., Marshall, D., Morton, H., and Jack, M. User perceptions of security and usability of single-factor and two-factor authentication in automated telephone banking. *Computers & Security* 30, 4 (2011), 208–220.

[87] Harbach, M., Zezschwitz, E. von, Fichtner, A., Luca, A. D., and Smith, M. It's a hard lock life: a field study of smartphone (un)locking behavior and risk perception. In: *Proc. 10th Symposium on Usable Privacy and Security (SOUPS'14).* USENIX Association, 2014.

[88] Harley, A. *Visibility of System Status.* https://www.nngroup.com/articles/visibility-system-status/. June 3, 2018.

[89] Hayashi, E. and Hong, J. A diary study of password usage in daily life. In: *Proc. CHI Conference on Human Factors in Computing Systems (CHI'11).* ACM, 2011.

[90] Herley, C., Oorschot, P. C. van, and Patrick, A. S. Passwords: if we're so smart, why are we still using them? In: *Proc. Financial Cryptography and Data Security (FC'09).* Springer, 2009.

[91] Herley, C. and Oorschot, P. van. A research agenda acknowledging the persistence of passwords. *IEEE Security and Privacy* 10, 1 (Jan. 2012), 28–36.

[92] Hitaj, B., Gasti, P., Ateniese, G., and Pérez-Cruz, F. Passgan: A deep learning approach for password guessing. *CoRR* abs/1709.00440 (2017). arXiv: `1709.00440`.

[93] *How To Enable and Use Password Generator in Google Chrome.* `https://edgetalk.net/enable-use-password-generator-google-chrome/`. Sept. 2016.

[94] Hox, J. J., Moerbeek, M., and Schoot, R. van de. *Multilevel Analysis: Techniques and Applications, Third Edition (Quantitative Methodology).* 3rd ed. An optional note. Quantitative Methodology Series, Sept. 2017.

[95] Huang, J. L., Bowling, N. A., Liu, M., and Li, Y. Detecting insufficient effort responding with an infrequency scale: evaluating validity and participant reactions. *Journal of Business and Psychology* 30, 2 (2015), 299–311.

[96] Hunt, T. *The science of password selection.* `https://www.troyhunt.com/science-of-password-selection/`. July 2011.

[97] Hunt, T. *The "Cobra Effect" that is disabling paste on password fields.* `https://www.troyhunt.com/the-cobra-effect-that-is-disabling/`. May 2014.

[98] Hunt, T. *It's not about "supporting password managers", it's about not consciously breaking security.* `https://www.troyhunt.com/its-not-about-supporting-password/`. July 2015.

[99] Hunt, T. *Password reuse, credential stuffing and another billion records in Have I been pwned.* `https://www.troyhunt.com/password-reuse-credential-stuffing-and-another-1-billion-records-in-have-i-been-pwned/`. May 2017.

[100] Inglesant, P. G. and Sasse, M. A. The true cost of unusable password policies: password use in the wild. In: *Proc. CHI Conference on Human Factors in Computing Systems (CHI'10).* ACM, 2010.

[101] *Interaction Design Foundation.* `https://www.interaction-design.org`.

[102] Interaction Design Foundation. *Gestalt Principles.* URL: `https://www.interaction-design.org/literature/topics/gestalt-principles`.

[103] International Organization for Standardization. *Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems.* Standard ISO 9241-210:2019. International Organization for Standardization, July 2019.

[104] Johnson, J. W. Factors affecting relative weights: the influence of sampling and measurement error. *Organizational Research Methods* 7, 3 (2004), 283–299.

[105] Karole, A., Saxena, N., and Christin, N. A comparative usability evaluation of traditional password managers. In: *Proc. 13th International Conference on Information Security and Cryptology.* Springer-Verlag, 2011.

[106] *Kaspersky Password Manager.* `https://www.kaspersky.com/password-manager`.

[107] Keil, M., Markert, P., and Dürmuth, M. “it’s just a lot of prerequisites”: a user perception and usability analysis of the german id card as a fido2 authenticator. In: *Proc. European Symposium on Usable Security (EuroUSEC ’22)*. ACM, 2022.

[108] Kelley, P. G., Komanduri, S., Mazurek, M. L., Shay, R., Vidas, T., Bauer, L., Christin, N., Cranor, L. F., and Lopez, J. Guess again (and again and again): measuring password strength by simulating password-cracking algorithms. In: *Proc. 33rd IEEE Symposium on Security and Privacy (SP ’12)*. IEEE, 2012.

[109] Koller, M. and Stahel, W. A. Sharpening wald-type inference in robust regression for small samples. *Computational Statistics & Data Analysis* 55, 8 (2011), 2504–2515.

[110] Koller, M. and Stahel, W. A. Nonsingular subsampling for regression s estimators with categorical predictors. *Computational Statistics* 32, 2 (June 2017), 631–646.

[111] Komanduri, S., Shay, R., Kelley, P. G., Mazurek, M. L., Bauer, L., Christin, N., Cranor, L. F., and Egelman, S. Of passwords and people: measuring the effect of password-composition policies. In: *Proc. CHI Conference on Human Factors in Computing Systems (CHI’11)*. ACM, 2011.

[112] Krause, R. *Maintain Consistency and Adhere to Standards (Usability Heuristic #4)*. https://www.nngroup.com/articles/consistency-and-standards/. Jan. 10, 2021.

[113] Krippendorff, K. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications, 2004.

[114] Krol, K., Philippou, E., Cristofaro, E. D., and Sasse, M. A. "they brought in the horrible key ring thing!" analysing the usability of two-factor authentication in uk online banking. In: *Workshop on Usable Security and Privacy (USEC’15)*. The Internet Society, 2015.

[115] Krug, S. *Don’t Make Me Think: A Common Sense Approach to Web Usability*. New Riders, 2013.

[116] Kumaraguru, P. and Cranor, L. F. *Privacy Indexes: A Survey of Westin’s Studies*. Tech. rep. 2005.

[117] Kunke, J., Wiefling, S., Ullmann, M., and Iacono, L. L. *Evaluation of Account Recovery Strategies with FIDO2-based Passwordless Authentication*. May 2021. arXiv: 2105.12477 [cs.CR].

[118] Lang, J., Czeskis, A., Balfanz, D., Schilder, M., and Srinivas, S. Security keys: practical cryptographic second factors for the modern web. In: *Proc. Financial Cryptography and Data Security (FC’17)*. 2017.

[119] Langer, M., König, C. J., and Fitili, A. Information as a double-edged sword: the role of computer experience and information on applicant reactions towards novel technologies for personnel selection. *Computers in Human Behavior* 81 (2018), 19–30.

[120] Lassak, L., Hildebrandt, A., Golla, M., and Ur, B. "it's stored, hopefully, on an encrypted server": mitigating users' misconceptions about fido2 biometric webauthn. In: *Proc. 30th USENIX Security Symposium (SEC' 21)*. USENIX Association, 2021.

[121] *LastPass*. `https://www.lastpass.com`.

[122] Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P., and Kort, J. Understanding, scoping and defining user experience: a survey approach. In: *Proc. CHI Conference on Human Factors in Computing Systems (CHI'09)*. ACM, 2009.

[123] Lazar, J. and Barbosa, S. D. J. Introduction to human-computer interaction. In: *Proc. CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, 2017.

[124] Lazar, J., Feng, J. H., and Hochheiser, H. *Research Methods in Human-Computer Interaction*. 2nd ed. Morgan Kaufmann, 2017.

[125] Le Pochat, V., Van Goethem, T., Tajalizadehkhoob, S., and Joosen, W. Tranco: a research-oriented top sites ranking hardened against manipulation. In: *Proc. 26th Annual Network and Distributed System Security Symposium (NDSS '19)*. The Internet Society, 2019.

[126] Lee, K., Sjöberg, S., and Narayanan, A. Password policies of most top websites fail to follow best practices. In: *Proc. 18th Symposium on Usable Privacy and Security (SOUPS'22)*. USENIX Association, 2022.

[127] Li, Z., He, W., Akhawe, D., and Song, D. The emperor's new password manager: security analysis of web-based password managers. In: *Proc. 23rd USENIX Security Symposium (SEC' 14)*. USENIX Association, 2014.

[128] M'Raihi, D., Machani, S., Pei, M., and Rydell, J. *TOTP: Time-Based One-Time Password Algorithm*. RFC 6238. `http://www.rfc-editor.org/rfc/rfc6238.txt`. RFC Editor, May 2011.

[129] Malhotra, N. K., Kim, S. S., and Agarwal, J. Internet users' information privacy concerns (iuipc): the construct, the scale, and a causal model. *Information systems research* 15, 4 (2004).

[130] Mardini, A. and Kim, G. *Making sign-in safer and more convenient*. `https://blog.google/technology/safety-security/making-sign-safer-and-more-convenient/`. Oct. 2021.

[131] McMillan, R. *The World's First Computer Password? It Was Useless Too*. `https://www.wired.com/2012/01/computer-password/`. 2012.

[132] Mehta, Y. *Windows Hello FIDO2 certification gets you closer to passwordless*. May 2019. URL: `https://techcommunity.microsoft.com/t5/Windows-IT-Pro-Blog/Windows-Hello-FIDO2-certification-gets-you-closer-to/ba-p/534592`.

[133] Melicher, W., Ur, B., Segreti, S. M., Komanduri, S., Bauer, L., Christin, N., and Cranor, L. F. Fast, lean, and accurate: modeling password guessability using neural networks. In: *Proc. 24th USENIX Security Symposium (SEC' 16)*. USENIX Association, 2016.

[134] Merriam, S. B. and Tisdell, E. J. *Qualitative research: A guide to design and implementation.* John Wiley & Sons, 2015.

[135] Microsoft. *Web Authentication and Windows Hello.* July 2018. URL: https://docs.microsoft.com/en-us/microsoft-edge/dev-guide/windows-integration/web-authentication.

[136] Milka, G. Anatomy of account takeover. In: *Enigma 2018.* USENIX Association, 2018.

[137] Mirian, A. Hack for hire. *Commun. ACM* 62, 12 (Nov. 2019), 32–37.

[138] Moore, P. *Don't let them paste passwords...* https://paul.reviews/dont-let-them-paste-passwords/. July 2015.

[139] Morris, R. and Thompson, K. Password security: a case history. *Commun. ACM* 22, 11 (Nov. 1979), 594–597.

[140] Napoli, D., Baig, K., Maqsood, S., and Chiasson, S. "i'm literally just hoping this will work:" obstacles blocking the online security and privacy of users with visual disabilities. In: *Proc. 17th Symposium on Usable Privacy and Security (SOUPS'21).* USENIX Association, 2021.

[141] Newman, L. H. *The new Yubikey will help kill the password.* Sept. 2018. URL: https://www.wired.com/story/yubikey-series-5-fido2-passwordless/.

[142] Nielsen, J. Enhancing the explanatory power of usability heuristics. In: *Proc. CHI Conference on Human Factors in Computing Systems (CHI'94).* ACM, 1994.

[143] Nielsen, J. *End of Web Design.* https://www.nngroup.com/articles/end-of-web-design/. July 22, 2000.

[144] Nielsen, J. *10 Usability Heuristics for User Interface Design.* https://www.nngroup.com/articles/ten-usability-heuristics/. Nov. 15, 2020.

[145] Nielsen, J. *Jakob's Law of Internet User Experience.* https://www.nngroup.com/videos/jakobs-law-internet-ux/.

[146] *Nielsen Norman Group.* https://www.nngroup.com/.

[147] Nielsen Norman Group. *Help and Documentation: The 10th Usability Heuristic.* https://www.nngroup.com/articles/help-and-documentation/.

[148] Norman, D. and Nielsen, J. *The Definition of User Experience (UX).* https://www.nngroup.com/articles/definition-user-experience/.

[149] Norman, D. A. and Draper, S. W. *User centered system design: new perspectives on human-computer interaction.* Lawrence Erlbaum Associates, Hillsdale, 1986.

[150] Notoatmodjo, G. and Thomborson, C. Passwords and perceptions. In: *Proc. 7th Australasian Conference on Information Security - Volume 98.* Australian Computer Society, Inc., 2009.

[151] Owens, K., Anise, O., Krauss, A., and Ur, B. User perceptions of the usability and security of smartphones as fido2 roaming authenticators. In: *Proc. 17th Symposium on Usable Privacy and Security (SOUPS'21).* USENIX Association, 2021.

[152] Parsons, K., Calic, D., Pattinson, M., Butavicius, M., McCormac, A., and Zwaans, T. The human aspects of information security questionnaire (HAIS-Q). *Comput. Secur.* 66, C (May 2017), 40–51.

[153] Parsovs, A. Practical issues with tls client certificate authentication. In: *Proc. 21th Annual Network and Distributed System Security Symposium (NDSS '14)*. The Internet Society, 2014.

[154] Pearman, S., Thomas, J., Naeini, P. E., Habib, H., Bauer, L., Christin, N., Cranor, L. F., Egelman, S., and Forget, A. Let's go in for a closer look: observing passwords in their natural habitat. In: *Proc. 24th ACM Conference on Computer and Communication Security (CCS '17)*. ACM, 2017.

[155] Peck, D. *WebAuthn and Biometrics*. Oct. 2018. URL: https://davepeck.org/2018/10/26/webauthn-and-biometrics/.

[156] Polybius. *The Histories of Polybius: Translated from the Text of F. Hultsch*. Ed. by Shuckburgh, E. S. Vol. 1. 2012.

[157] Powers, A. *A node.js library for performing FIDO 2.0 / WebAuthn server functionality*. URL: https://github.com/apowers313/fido2-lib.

[158] Powers, A. *A simple WebAuthn / FIDO2 JavaScript application*. URL: https://github.com/apowers313/webauthn-simple-app.

[159] Ranger, S. *Windows 10: We're going to kill off passwords and here's how, says Microsoft*. May 2018. URL: https://www.zdnet.com/article/windows-10-were-going-to-kill-off-passwords-and-heres-how-says-microsoft/.

[160] Redmiles, E. M., Warford, N., Jayanti, A., Koneru, A., Kross, S., Morales, M., Stevens, R., and Mazurek, M. L. A comprehensive quality evaluation of security and privacy advice on the web. In: *Proc. 29th USENIX Security Symposium (SEC' 20)*. USENIX Association, 2020.

[161] Reese, K., Smith, T., Dutson, J., Armknecht, J., Cameron, J., and Seamons, K. A usability study of five two-factor authentication methods. In: *Proc. 15th Symposium on Usable Privacy and Security (SOUPS'19)*. USENIX Association, 2019.

[162] Rescorla, E. *The Transport Layer Security (TLS) Protocol Version 1.3*. RFC 8446. Aug. 2018.

[163] Reynolds, J., Samarin, N., Barnes, J., Judd, T., Mason, J., Bailey, M., and Egelman, S. Empirical measurement of systemic 2fa usability. In: *Proc. 29th USENIX Security Symposium (SEC' 20)*. USENIX Association, Aug. 2020.

[164] Reynolds, J., Smith, T., Reese, K., Dickinson, L., Ruoti, S., and Seamons, K. A tale of two studies: the best and worst of yubikey usability. In: *Proc. 39th IEEE Symposium on Security and Privacy (SP '18)*. IEEE, 2018.

[165] Rinn, C., Summers, K., Rhodes, E., Virothaisakun, J., and Chisnell, D. Password creation strategies across high- and low-literacy web users. In: *Proc. 78th ASIS&T Annual Meeting (ASIST'15)*. American Society for Information Science, 2015.

[166] Rosala, M. *User Control and Freedom.* https://www.nngroup.com/articles/user-control-and-freedom/. Nov. 20, 2020.

[167] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20 (1987), 53–65.

[168] Rubenking, N. J. *The Best Password Managers of 2017.* http://uk.pcmag.com/password-managers-products/4296/guide/the-best-password-managers-of-2017. Nov. 2017.

[169] Sauro, J. *A practical guide to the system usability scale: Background, benchmarks & best practices.* Measuring Usability LLC Denver, CO, 2011.

[170] Schwarz, N. and Sudman, S. *Context effects in social and psychological research.* Springer Science & Business Media, 2012.

[171] SciPy Developer Documentation. *Statistical functions: normaltest.* https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html.

[172] Shay, R., Komanduri, S., Kelley, P. G., Leon, P. G., Mazurek, M. L., Bauer, L., Christin, N., and Cranor, L. F. Encountering stronger password requirements: user attitudes and behaviors. In: *Proc. 6th Symposium on Usable Privacy and Security (SOUPS'10).* ACM, 2010.

[173] Shneiderman, B. *Designing the user interface : strategies for effective human-computer interaction.* 4th ed. Pearson/Addison Wesley, Boston, 2004.

[174] Shneiderman, B. *The Eight Golden Rules of Interface Design.* https://www.cs.umd.edu/~ben/goldenrules.html.

[175] Silver, D., Jana, S., Boneh, D., Chen, E., and Jackson, C. Password managers: attacks and defenses. In: *Proc. 23rd USENIX Security Symposium (SEC' 14).* USENIX Association, 2014.

[176] Simons, A. *Secure password-less sign-in for your Microsoft account using a security key or Windows Hello.* Nov. 2018. URL: https://www.microsoft.com/en-us/microsoft-365/blog/2018/11/20/sign-in-to-your-microsoft-account-without-a-password-using-windows-hello-or-a-security-key/.

[177] Stajano, F. Pico: no more passwords! In: *Security Protocols Workshop.* Springer, 2011.

[178] Stobert, E. and Biddle, R. The password life cycle: user behaviour in managing passwords. In: *Proc. 10th Symposium on Usable Privacy and Security (SOUPS'14).* USENIX Association, 2014.

[179] Stock, B. and Johns, M. Protecting users against xss-based password manager abuse. In: *Proc. 9th ACM Symposium on Information, Computer and Communication Security (ASIACCS '14).* ACM, 2014.

[180] Strouble, D., Shechtman, G. m., and Alsop, A. S. Productivity and usability effects of using a two-factor security system. In: *SAIS.* 2009.

[181] Tam, L., Glassman, M., and Vandenwauver, M. The psychology of password management: a tradeoff between security and convenience. *Behaviour & Information Technology* 29, 3 (2010), 233–244.

[182] Tetlay, A., Treharne, H., Ascroft, T., and Moschoyiannis, S. Lessons learnt from a 2fa roll out within a higher education organisation. *CoRR* abs/2011.02901 (2020). eprint: 2011.02901.

[183] The University of Chicago – IT Services. *Strengthen Your Passwords or Passphrases and Keep Them Secure.* https://uchicago.service-now.com/it?id=kb_article&kb=KB00015347. Oct. 2017.

[184] Tranco. *A Research-Oriented Top Sites Ranking Hardened Against Manipulation.* https://tranco-list.eu/.

[185] Trewin, S., Swart, C., Koved, L., Martino, J., Singh, K., and Ben-David, S. Biometric authentication on a mobile device: a study of user effort, error and task disruption. In: *Proc. 28th Annual Computer Security Applications Conference (ACSAC'12).* ACM, 2012.

[186] U.S. Census Bureau. *2010 Census National Summary File of Redistricting Data.* https://www.census.gov/2010census/data/. 2011.

[187] *United States v. Morris (1991), 928 F.2d 504, 505 (2d Cir.)* https://scholar.google.com/scholar_case?case=551386241451639668. Mar. 1991.

[188] Ur, B., Alfieri, F., Aung, M., Bauer, L., Christin, N., Colnago, J., Cranor, L. F., Dixon, H., Naeini, P. E., Habib, H., Johnson, N., and Melicher, W. Design and evaluation of a data-driven password meter. In: *Proc. CHI Conference on Human Factors in Computing Systems (CHI'17).* ACM, 2017.

[189] Ur, B., Kelley, P. G., Komanduri, S., Lee, J., Maass, M., Mazurek, M. L., Passaro, T., Shay, R., Vidas, T., Bauer, L., Christin, N., and Cranor, L. F. How does your password measure up? the effect of strength meters on password creation. In: *Proc. 21st USENIX Security Symposium (SEC '12).* USENIX Association, 2012.

[190] Ur, B., Noma, F., Bees, J., Segreti, S. M., Shay, R., Bauer, L., Christin, N., and Cranor, L. F. "i added '!' at the end to make it secure": observing password creation in the lab. In: *Proc. 11th Symposium on Usable Privacy and Security (SOUPS'15).* USENIX Association, 2015.

[191] Ur, B., Segreti, S. M., Bauer, L., Christin, N., Cranor, L. F., Komanduri, S., Kurilova, D., Mazurek, M. L., Melicher, W., and Shay, R. Measuring real-world accuracies and biases in modeling password guessability. In: *Proc. 24th USENIX Security Symposium (SEC' 15).* USENIX Association, 2015.

[192] Vaas, L. *Android nudges passwords closer to the cliff edge with FIDO2 support.* Feb. 2019. URL: https://nakedsecurity.sophos.com/2019/02/26/android-nudges-passwords-closer-to-the-cliff-edge-with-fido2-support/.

[193] Venkatesh, V. and Davis, F. D. A theoretical extension of the technology acceptance model: four longitudinal field studies. *Management Science* 46, 2 (2000), 186–204.

[194] *W3Counter: Browser & Platform Market Share (November 2017)*. https://www.w3counter.com/globalstats.php.

[195] Wang, D. and Wang, P. The emperor's new password creation policies: an evaluation of leading web services and the effect of role in resisting against online guessing. In: *Proc. 20th European Symposium on Research in Computer Security (ESORICS'15)*. Springer, 2015.

[196] Wash, R., Rader, E., Berman, R., and Wellmer, Z. Understanding password choices: how frequently entered passwords are re-used across websites. In: *Proc. 12th Symposium on Usable Privacy and Security (SOUPS'16)*. USENIX Association, 2016.

[197] Weidman, J. and Grossklags, J. I like it, but i hate it: employee perceptions towards an institutional transition to byod second-factor authentication. In: *Proc. 33rd Annual Computer Security Applications Conference (ACSAC'17)*. ACM, 2017.

[198] Weiler, J. *Two-Factor Authentication – better user experience through security.* https://amiconsult.de/en/better-user-experience-through-security/. Dec. 2020.

[199] Weinert, A. *Your Pa$$word doesn't matter.* https://techcommunity.microsoft.com/t5/azure-active-directory-identity/your-pa-word-doesn-t-matter/ba-p/731984. July 2019.

[200] Weinschenk, S. *100 Things Every Designer Needs to Know about People.* 2nd. New Riders, 2020.

[201] Weir, C. S., Douglas, G., Carruthers, M., and Jack, M. User perceptions of security, convenience and usability for ebanking authentication tokens. *Computers & Security* 28, 1 (2009), 47–62.

[202] Weir, C. S., Douglas, G., Richardson, T., and Jack, M. Usable security: user preferences for authentication methods in ebanking and the effects of experience. *Interacting with Computers* 22, 3 (2010), 153–164.

[203] Weir, M., Aggarwal, S., Collins, M., and Stern, H. Testing metrics for password creation policies by attacking large sets of revealed passwords. In: *Proc. 17th ACM Conference on Computer and Communication Security (CCS '10)*. ACM, 2010.

[204] West, R. The psychology of security. *Commun. ACM* 51, 4 (Apr. 2008), 34–40.

[205] Wheeler, D. L. Zxcvbn: low-budget password strength estimation. In: *Proc. 24th USENIX Security Symposium (SEC' 16)*. USENIX Association, 2016.

[206] Wiefling, S., Lo Iacono, L., and Dürmuth, M. Is This Really You? An Empirical Study on Risk-Based Authentication Applied in the Wild. In: *34th IFIP TC-11 International Conference on Information Security and Privacy Protection (IFIP SEC 2019)*. Springer International Publishing, 2019.

[207] Wielgoszewski, M. *Securing your Gemini Account with WebAuthn.* May 2019. URL: https://medium.com/gemini/securing-your-gemini-account-with-webauthn-b5f369b8beec.

[208] Wilhelmy, A. Journal guidelines for qualitative research? a balancing act that might be worth it. *Industrial and Organizational Psychology* 9, 4 (2016), 726–732.

[209] Wolf, F., Kuber, R., and Aviv, A. J. "pretty close to a must-have": balancing usability desire and security concern in biometric adoption. In: *Proc. CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, 2019.

[210] Wong, B. *WebAuthn: The future of device based 2FA at Twitter*. May 2019. URL: https://blog.twitter.com/engineering/en_us/topics/infrastructure/2019/webauthn.html.

[211] Woodruff, A., Pihur, V., Consolvo, S., Brandimarte, L., and Acquisti, A. Would a privacy fundamentalist sell their DNA for $1000...if nothing bad happened as a result? the westin categories, behavioral intentions, and consequences. In: *Proc. 10th Symposium on Usable Privacy and Security (SOUPS'14)*. USENIX Association, 2014.

[212] World Wide Web Consortium. *Web Authentication: An API for accessing Public Key Credentials Level 1 — W3C Recommendation, 4 March 2019*. Mar. 2019. URL: https://www.w3.org/TR/webauthn/.

[213] World Wide Web Consortium. *Web Authentication: An API for accessing Public Key Credentials Level 2 — W3C Recommendation, 8 April 2021*. https://www.w3.org/TR/webauthn/. Mar. 2021.

[214] Yablonski, J. *Laws of UX: Design Principles for Persuasive and Ethical Products*. O'Reilly UK Ltd., 2020.

[215] Yablonski, J. *Laws of UX*. https://lawsofux.com/. 2021.

[216] Yadron, D. *Man Behind the First Computer Password: It's Become a Nightmare*. https://www.wsj.com/articles/BL-DGB-35227. May 2014.

[217] Yan, J., Blackwell, A., Anderson, R., and Grant, A. Password memorability and security: empirical results. *IEEE Security and Privacy* 2, 5 (Sept. 2004), 25–31.

[218] Yubico. *Developer Program*. 2019. URL: https://developers.yubico.com.

[219] Yubico Developer Program. *Card edit*. URL: https://developers.yubico.com/PGP/Card_edit.html.

[220] Zhao, R., Yue, C., and Sun, K. Vulnerability and risk analysis of two commercial browser and cloud based password managers. *ASE Science Journal* 1, 4 (2013), 1–15.

A
Appendix: Study of Password Manager Usage

A.1 Sampling Survey Questions

Q1: For each of the following statements, how strongly do you agree or disagree?

a1: Consumer have lost all control over how personal information is collected and used by companies.

a2: Most businesses handle the personal information they collect about consumers in a proper and confidential way.

a3: Existing laws and organizational practices provide a reasonable level of protection for consumer privacy today.

(i) Strongly disagree, (ii) Somewhat disagree, (iii) Somewhat agree, (iv) Strongly agree

Q2: On how many different Internet sites do you have a user account that is secured with a password? (If you are not sure about the number please estimate the number) (FreeText)

Q3: Has ever one of your passwords been leaked or been stolen?

(i) Yes, (ii) No, (iii) I am not aware of that, (iv) I do not care

Q4: How strongly do you agree or disagree?

b1. Passwords are useless, because hackers can steal my data either way. (i) Strongly disagree, (ii) Somewhat disagree, (iii) Somewhat agree, (iv) Strongly agree

b2. I don't care about my passwords' strength, because I don't have anything to hide.

(i) Strongly disagree, (ii) Somewhat disagree, (iii) Somewhat agree, (iv) Strongly agree

Q5: What characterizes in your opinion a strong/secure password? (FreeText)

Q6: Please rate the strength of the following passwords?

c1. thHisiSaSecUrePassWord

c2. Pa$sWordsk123

c3. AiWuutaiveep9j

c4. !@#$%&̂*()

c5. 12/07/2017

(i) Very weak, (ii) Weak, (iii) Moderate strength, (iv) Strong, (v) Very strong

Q7: I have never used a computer? (i) I have never, (ii) I do

Q8: How would you rate your ability to create strong passwords?

(i) 5 (high ability), (ii) 4, (iii) 3, (iv) 2, (v) 1 (low ability)

Q9: How do you proceed if you have to create a new password? (What is your strategy?) (FreeText)

Q10: I try to create secure passwords.....

(i) for all my accounts and websites, (ii) for my email accounts, (iii) for online shopping, (iv) for online booking/reservation, (v) for social networks, (vi) No answer, (vii) Other

Q11: I make a point of changing my passwords on websites that are critical to my privacy every...... (choose the closest match)
(i) Day, (ii) Week, (iii) Two weeks, (iv) Month, (v) 6 month, (vi) Year, (vii) Never, (viii) Other

Q12: Do you use the same password for different email accounts, websites, or devices? (i) Yes, (ii) No

Q13: Do you use any of the following strategies for creating your password or part of your password, anywhere, at any time in the last year... (i) I used the name of celebrities as a password or as a part of a password, (ii) I used the name of family members as a password or as a part of a password, (iii) I used literature (book, poetry, etc.) as a password or as a part of a password, (iv) I used familiar numbers (street address, employee number, etc) as a password or as a part of a password, (v) I used random characters as a password, (vi) I used a password manager to generate passwords, (vii) No answer, (viii) Other

Q14: How do you remember all of your passwords? (i) I write them down on paper (notebook, day planner, etc), (ii) I try to remember them (human memory), (iii) I use computer files (Word document, Excel sheet, text file, etc), (iv) I use encrypted computer files (e.g. CryptoPad), (v) I store my passwords on my mobile phone or PDA, (vi) I use 3rd party password manager (save in extra program, e.g. LastPass, keepass, 1Password, etc.), (vii) I use website cookies (Website checkbox: “Remember my password on this computer”), (viii) I use the same password for more than one purpose, (ix) I use browser built-in password manager (i.e saved in browser), (x) I use a variation of a past password (eg. password1 and then password2 and then password3, etc.), (xi) No answer, (xii) Other

Q15: Have you ever used a computer program to generate your passwords? (i) Yes, (ii) No

Q16: When creating a new password, which do you regard as most important: choosing a password that is easy to remember for future use (ease of remembering) or the password’s security?
(i) Always ease of remembering, (ii) Mostly ease of remembering, (iii) Mostly security, (iv) Always security, (v) Other

Q17: When you create a new password, which of the following factors do you consider? The password ....
(i) does not contain dictionary words, (ii) is in a foreign (non-English) language, (iii) is not related to the site (i.e., the name of the site), (iv) includes numbers, (v) includes special characters (e.g. “&” or “!”), (vi) is at least eight (8) characters long, (vii) None of the above: I didn’t think about it, (viii) No answer, (ix) Other

Q18: My home planet is Earth? (i) Yes, (ii) No

Q19: Do you use the "save password" feature of your browser? (i) Yes, (ii) No

Q20: Do you use any kind of extra password manager program (for instance, LastPass, 1Password, Keepass, Dashlane, etc.)? (i) Yes, (ii) No

Q21: Which password manager(s) do you use? (You can write one name per line) (FreeText)

Q22: Please give us a short description of your impression of using your browser's password saving feature and/or of using extra password managers (FreeText)

Q23: How many passwords do you keep in your password manager(s) and browser's saved passwords? (if you don't know the exact number, please estimate the number) (FreeText)

A.2 Zxcvbn Score

To better understand zxcvbn's scoring, we used zxcvbn to score 200 million unique passwords collected from `hashes.org`, where we measured the zxcvbn score and the corresponding guesses in log10. The results in Table A.1 show that each score has a corresponding cutoff for guesses, e.g., score 2 requires between 10^3–10^6 guesses.

Score	#Passwords	Mean	SD	Min	25%	50%	75%	Max
0	122,296	2.69	0.42	0.30	2.48	2.92	3.00	3.00
1	34,496,960	5.34	0.59	3.00	5.00	5.44	5.87	6.00
2	69,090,776	7.15	0.66	6.00	6.61	7.00	7.87	8.00
3	57,256,840	8.87	0.65	8.00	8.28	8.87	9.36	10.00
4	39,789,207	12.51	2.29	10.00	11.00	12.00	13.36	32.00

Table A.1: Zxcvbn scores and estimated no. of guesses (in log10) for 200 million unique passwords from *hashes.org*.

A.3 Website category vs. website value

Commonly the website category is used as a proxy for the website value. Since we collected both, we can provide insights into this general assumption. Figure A.1 shows the self-reported value per domain. For instance, in >70% of logged passwords for a financial domain, the user reported a very high value for that domain. Similarly, in >60% of a logged passwords for news websites, the users (strongly) disagreed that this domain has a high value.

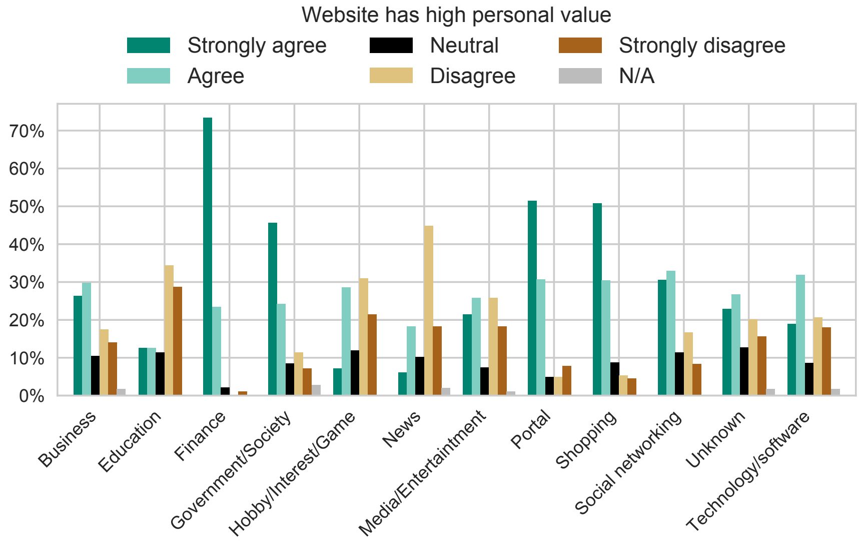


Figure A.1: Self-reported website value per website category.

A.4 Known Password Manager Plugins

Chrome plugins are identified through a 32 characters long UUID that can be retrieved from Google's Chrome Web Store. Table A.2 lists the password manager plugins that our study plugin can detect based on their UUID. Plugins not in this list are reported as *"Unknown plugin."*

Name	**UUID**
Dashlane	fdjamakpfbbddfjaooikfcpapjohcfmg
LastPass	hdokiejnpimakedhajhdlcegeplioahd
1Password	aomjjhallfgjeglblehebfpbcfeobpgk
Roboform	pnlccmojcmeohlpggmfnbbiapkmbliob
Enpass	kmcfomidfpdkfieipokbalgegidffkal
Zoho Vault	igkpcodhieompeloncfnbekccinhapdb
Norton Identity Safe	iikflkcanblccfahdhdonehdalibjnif
KeePass	ompiailgknfdndiefoaoiligalphfdae

Table A.2: UUIDs of plugins detected by our study plugin.

A.5 Model fit

All models in the building process were compared according to the corresponding akaike information criterion (AIC), which is an estimator of the relative quality of statistical models for a given set of data. Additionally, the models were statistically compared using likelihood-ratio tests, which were evaluated using a Chi-squared distribution. The final model is selected based on AIC as well as their ability to describe the empirical data better than the previous models. Tables A.3 and A.4 present the goodness of fit for the relevant steps in the model building process.

	AIC	logLik	df	Pr(>Chisq)
simple regression	5080.6	-2536.3		
multi-level base	4536.7	-2263.4	1	<0.001
+ login level	4316.3	-2147.1	6	<0.001
+ user level	4320.4	-2143.2	6	0.2494034
+ interactions	4309.5	-2133.7	4	<0.001

Table A.3: Goodness of fit for models predicting ZCVBN scores.

	AIC	logLik	Df	Pr(>Chisq)
simple regression	1959.7	-978.84		
multi-level base	1794.6	-895.28	1	< 0.001
+ login level	1694.9	-839.46	6	< 0.001
+ user level	1684.7	-828.37	6	<0.01
+ interactions	1687.6	-825.80	4	0.27351

Table A.4: Goodness of fit for models predicting password reuse.

B
Appendix: Study of FIDO2/WebAuthn Passwordless Authentication

B.1 Survey Questions

Acceptance: Please judge the presented authentication method on the following adjectives.

Useless	○	○	○	○	○	Useful
Unpleasant	○	○	○	○	○	Pleasant
Bad	○	○	○	○	○	Good
Annoying	○	○	○	○	○	Nice
Superfluous	○	○	○	○	○	Effective
Irritating	○	○	○	○	○	Likeable
Worthless	○	○	○	○	○	Assisting
Undesirable	○	○	○	○	○	Desirable
Sleep-inducing	○	○	○	○	○	Raising alertness

System Usability Scale (SUS): Please state your level of agreement or disagreement for the following statements based on your experience with the presented authentication method. There are no right or wrong answers. (Strongly disagree; Disagree; Neither disagree nor agree; Agree; Strongly agree.)

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very awkward to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Affinity for Technology Interaction (ATI): In the following, we will ask you about your interaction with technical systems. The term “technical systems” refers to apps and other software applications, as well as entire digital devices (e.g., Mobile phone, computer, TV, car navigation). Please indicate the degree to which you agree/disagree with the following statements. There are no right or wrong answers. (Completely disagree; Largely disagree; Slightly disagree; Slightly agree; Largely agree; Completely agree)

1. I like to occupy myself in greater detail with technical systems.
2. I like testing the functions of new technical systems.
3. I predominantly deal with technical systems because I have to.

4. When I have a new technical system in front of me, I try it out intensively.
5. I enjoy spending time becoming acquainted with a new technical system.
6. It is enough for me that a technical system works; I don't care how or why.
7. I try to understand how a technical system exactly works.
8. It is enough for me to know the basic functions of a technical system.
9. I try to make full use of the capabilities of a technical system.

Privacy Concern: Please state how much you agree or disagree to the following statements. There are no right or wrong answers. (Strongly disagree; Disagree; somewhat disagree; Neither disagree nor agree; somewhat agree; Agree; Strongly agree.)

1. I am concerned that companies are collecting too much information about me.
2. I am concerned about my privacy.
3. To me it is important to keep my privacy intact.
4. Novel technologies are threatening privacy increasingly.

Technical Problems: Were there any technical problems while watching the video and trying out the presented authentication method?
(i) No problems, (ii) Few problems, (iii) Some problems, (iv) Many problems
If you have experienced technical problems before, during, or after watching the video or trying out the presented authentication method, please describe them briefly.
Open-Ended Questions: How would you describe your general experience about the presented authentication with a Yubikey[1] (Only in Group$_{\text{1FA}}$)? [Free response]
Which advantages do you see in the usage of the presented authentication method ? [Free response]
Which disadvantages do you see in the usage of the presented authentication method? [Free response]
Would you use the presented authentication method yourself? If you would, why and on which accounts would you use it? If you wouldn't, why not (Only in Group$_{\text{1FA}}$)? [Free response]
Further Questions: How do you choose your password for a new email account? (i) Reuse an existing password, (ii) Modify an existing password, (iii) Create an entirely new password, (iv) No answer, (v) Other
Has ever one of your passwords been leaked or been stolen? (i) yes, (ii) No
Demographic Questions: Please indicate your gender.
(i) Male, (ii) Female, (iii) Other, (iv) No answer
Please indicate your highest educational degree. (i) High school graduate, (ii) Bachelor's degree, (iii) Master's degree, (iv) Diploma, (v) Ph.D, (vi) Other
How old (in years) are you? Free response
Please indicate if you have a computer science background. (i) Yes, (ii) No
Please indicate your area of studies/area of work. Free response

[1] For simplicity for our participants, we called the "Yubico Security Key" just "Yubikey" in our study

B.2 Model Comparison

	Res.Df	R^2adj.	step-wise comparison		
			δDf	W	p
Users characteristics	90	<0.1%			
+ usability	89	28.0%	1	35.96	**<.01**
+ authentication type	88	48.8%	1	34.47	**<.01**
+ interactions	84	51.5%	4	7.26	.12

Note: Res.Df = Residual Degrees of freedom, R^2adj. = Percentage of the empirical variance that could be explained by the regression model (adjusted for number of terms in model). Model 3 explains the empirical data best under the conditions of parsimony (Occam's razor), δdf = differences in the number of constraints between models, W = Wald statistic, p values below the 5% criterion are printed in bold. N(total) = 94.

Table B.1: Model Comparison.

B.3 Results including Group$_{\text{1FAcon}}$

Variable	Group			Statistics	ES
	Pass	1FA	1FAcon		
N	48	46	47		
Gender				$\chi^2(2) = 1.923$	.12
Female	27	26	20	$p = .392$	
Male	20	20	25		
No answer	1	0	2		
Age	24.08	25.78	25.21	$F(2, 138) = 1.479$	.02
	(3.63)	(6.44)	(4.19)	$p = .231$	
Education				$\chi^2(8) = 14.462$	.23
< High school	0	2	3	$\boldsymbol{p = .026}$	
High school	23	12	10		
Bachelor	12	20	25		
Master	12	11	8		
Diploma	0	1	0		
Ph.D	1	0	1		
ATI	3.84	4.01	4.05	$F(2) = 0.533$	.01
	(1.12)	(0.95)	(1.03)	$p = .588$	
PC	5.43	5.36	5.63	$F(2, 138) = 0.668$	.00
	(1.31)	(1.13)	(1.07)	$p = .514$	
CS background				$\chi^2(2) =6.047$	.21
Yes	18	28	27	$\boldsymbol{p = .047}$	
No	30	18	20		
SUS	71.92	81.79	79.20	$F(2, 138) = 9.122$	.12
	(11.09)	(12.15)	(11.91)	$\boldsymbol{p < .001}$	
Acceptance	3.41	4.29	4.16	$F(2, 138) = 24.420$	.26
	(0.70)	(0.60)	(0.66)	$\boldsymbol{p < .001}$	

Note: ES = Effect Size; N = Number of participants; < High school = Less than high school; ATI = Affinity for Technology Interaction; PC = Privacy Concerns; CS background = Computer science background; SUS = System Usability Scale. Depending on the variable, the frequencies or the scale mean values including standard deviation are presented in the cells. The statistics column shows the statistical data parameters for a group comparison with one-way anova respectively with Fisher's exact test for the corresponding variable. p values below the 5% criterion are printed in bold. Effect Sizes are specified in Eta-squared (η^2) for one-way anova and in Cramer's V for Fisher's exact test. N(total) = 141.

Table B.2: Overview descriptive data including Group$_{\text{1FAcon}}$.

	Acceptance			
Predictors	b	CI	RI	p
(Intercept)	3.57	[3.38 , 3.75]		**<0.001**
ATI	−0.01	[−0.12 , 0.10]	0.9%	0.846
PC	−0.02	[−0.10 , 0.06]	0.3%	0.591
CS (yes)	−0.12	[−0.34 , 0.10]	0.7%	0.269
SUS	0.03	[0.02 , 0.04]	56.2%	**<0.001**
Group			41.9%	
1FA	0.70	[0.47 , 0.94]		**<0.001**
1FAcon	0.66	[0.42 , 0.89]		**<0.001**

Note: Robust regression based on MM estimator [109]. Model 3 can explain 47.1% (R^2adjusted = .471) of the empirical variance (adjusted for number of terms in model); ATI = Affinity for Technology Interaction; PC = Privacy Concerns; CS (yes) = Dummy variable that encodes the effect of a computer science background (No background is the default); SUS = System Usability Scale; $\text{Group}_{\text{1FA}}$ = Dummy variable that encodes the differences for the groups ($\text{Group}_{\text{Pass}}$ is the default). $\text{Group}_{\text{1FAcon}}$ = Dummy variable that encodes the differences for the groups ($\text{Group}_{\text{Pass}}$ is the default). p-values below the 5% criterion are printed in bold. N(total) = 141.

Table B.3: Regression model predicting users acceptance data including $\text{Group}_{\text{1FAcon}}$.

	Res.Df	R^2adj.	step-wise comparison		
			δDf	W	p
Users characteristics	137	<0.1%			
+ usability	136	31.0%	1	61.15	**<.01**
+ authentication type	134	47.1%	2	40.44	**<.01**
+ interactions	126	49.9%	8	15.39	.06

Note: Res.Df = Residual Degrees of freedom, R^2adj. = Percentage of the empirical variance that could be explained by the regression model (adjusted for number of terms in model). Model 3 explains the empirical data best under the conditions of parsimony (Occam's razor), δdf = differences in the number of constraints between models, W = Wald statistic, p values below the 5% criterion are printed in bold. N(total) = 141.

Table B.4: Model Comparison including $\text{Group}_{\text{1FAcon}}$.

Category	$\textbf{Group}_{\textbf{1FA}}$	$\textbf{Group}_{\textbf{1FAcon}}$
N	46	47
Mental effort		
Reduction of cognitive effort	30 (65%)	37 (79%)
Efficient and easy to use	35 (76%)	37 (79%)
Physical Effort		
Carrying an extra device	16 (35%)	14 (30%)
Threat model		
Device theft/loss	28 (61%)	36 (77%)
Access to account by owner (recovery)	11 (24%)	7 (15%)
Access to account by other (revocation)	8 (17%)	23 (49%)
Fallback authentication	12 (26%)	7 (15%)
Applicability		
Device and connectivity support	14 (30%)	6 (13%)
System transperancy		
Mistrust	9 (20%)	22 (47%)
Lack of knowledge	8 (17%)	22 (47%)
Perceived security	20 (44%)	12 (26%)
Affective perception		
Fun / Excitement	22 (48%)	18 (38%)
Security Key characteristics		
Robustness and maturity	7 (15%)	10 (21%)
Cost	10 (22%)	2 (4%)

Note: (N): Number of participants in both $\text{Group}_{\text{1FA}}$ and $\text{Group}_{\text{1FAcon}}$. Categories that are mentioned here are based on the code book in Table 3.2

Table B.5: Comparison qualitative data.

C

Appendix: Study of Consistency of Two-Factor Authentication User Journeys

C.1 Sampling Survey about 2FA User Experience

We conducted a survey among 2FA users to gather users' experiences with different 2FA journeys and gain insights into whether users had negative experiences transferring their 2FA experiences between websites and whether this has stopped them from enabling or continuing to use 2FA. In the following, we first explain the structure of our survey (Appendix C.1.1), followed by an overview of the results (Appendix C.1.2, and lastly we provide a brief conclusion (Appendix C.1.3).

C.1.1 Sampling Survey Questions

The structure of our survey was as follows: First, we presented a general welcome message that explained 1) that this survey is about 2FA user experiences, 2) which types of personal data are collected and how the collected data is handled, 3) that participation is voluntary and participants have the right to withdraw their consent at any time during the study, 4) that are bonus tasks in this survey but not when and how, and 5) who is conducting the study and which contact points exist. Afterward, participants followed the survey structure below, on which we converged after three rounds of pre-testing with 10 randomly selected Prolific participants in each round (cf. Appendix C.1.2.1):

[**Q1: Same notion of 2FA**] As you mentioned in the first part of the study that you use two-factor authentication on different websites, we assume that we don't need to explain the concept to you anymore. However, to make sure that we all refer to exactly the same thing, here is a short definition from our side: Two-factor authentication (short: 2FA, but also sometimes referred to as two-step verification) is a method to confirm a user has claimed online identity by using a combination of two different types of authentication methods. Typically a password is considered one type, and with 2FA, the password is combined with another type to increase login security (for example, code via SMS or from an app like Google Authenticator or Duo Security, or fingerprint). However, websites often differ in how they implement 2FA in detail and therefore also in the experience for the user.

Does this definition match what you mean by 2FA? (i) Yes, (ii) No

[**Q2: Current and past use of 2FA**] Please tell us on which websites you currently or in the past have used 2FA. Additionally, we are also interested in websites where you have started to set up 2FA but have aborted the setup (please scroll further down to enter those).

[**Q2a**] *I currently use 2FA on the following websites: (Up to 20, please name the site and not use generic terms such as "banking websites"):* [**Form field**]

[**Q2b**] *I have used 2FA on the following websites in the past but stopped using it, or I started to set up 2FA on these websites but aborted the setup (up to 10, please name the site and not use generic terms such as "banking websites"):* [**Form field**]

[**Q3: Recognized differences between websites**] Please read the following instructions really carefully and take your time to think about your answer! Since this question is really important to us, we have set a short timer so that you can continue the survey only after 45 seconds (the "Next" button will appear automatically). Now, please take a moment to compare your 2FA experiences with all websites you listed on the previous page *(they are also listed below on this page)*. Those experiences include how the individual sites present 2FA, the options for the second factor each site offers, or the steps to set up 2FA and log in on each site. Please note that we are not asking about your general impression of 2FA but the differences between the 2FA on different websites that you noticed. If you did not notice such differences, please answer accordingly below.

Are there one or more websites where your experiences, based on the above definition, differed significantly positively or negatively from the other sites? If so, select them below. If you did not notice any differences, select none of the websites and click "Next". [Website A], [Website B]...[Website N]

[**If no website selected in Q3 go to Q6**]

[**Q4: Opt-in to bonus task**] You are almost through the survey, and only one page of questions about your experience with 2FA is missing. However, we would like to offer you a bonus task where you can earn some extra money. On the previous page, you stated that your experience with 2FA is very different on the following website(s): [website A] [Website B] ... [Website N]. We would be very interested in what exactly made the difference in your experiences with 2FA on these websites. For each of these sites, we have a few questions for you. It usually takes no longer than 2 minutes to answer all the questions per website, and you will get paid a bonus of 50 cents per website. With this bonus task, you can earn another [number of websites in Q3 × 0.50] pounds with this bonus task.
Would you like to do the bonus task?
(i) Yes, take me to the bonus task!, (ii) No, take me to the last questions of your survey!

[**If answer to Q4 is *No* then go to Q6**]

[**For each website selected in Q3 ask:**]

> [**Q5**] Please tell us a little bit more about [Website].
> [**Q5a: Differences**] *How exactly did your 2FA experience differ on [Website] from the other websites that you listed? (Those experiences include how [Website] presents 2FA, its options for the second factor, or the steps to set up 2FA and log in on [Website])[**Open-text question**]*
> [**Q5b: Usability**] *Did this difference make 2FA on [Website] easier or harder for you to use in comparison to other websites with 2FA? (i) Easier, (ii) Equally easy, (iii) harder*
> [**Q5c: Behavior**] *How has this changed your usage behavior on [Website] compared to other websites on which you use 2FA? For example, do you log in to the website more/less? Or was that maybe even a reason why you don't use 2FA on the site in the end?* [**Open-text question**]
> [**Q5d: Recommendation**] *Based on your experience of 2FA on [Website] Is there something that the other websites should adopt for their 2FA or should they avoid?* [**Open-text question**]

[**Q6: Concrete situation**] You might be used to a particular experience of 2FA, meaning how websites present 2FA, which second factors you can use, or how to set 2FA up and log in.
[**Q6a: Problematic inconsistency**] *Do you remember a concrete situation where your 2FA experience was challenging to you because it differed from what you were used to? Please give us as many details as you can remember about this situation. [**Open-text question**].*
[**Q6b: Solution**] *How did you behave in this situation? For example, you aborted setting up 2FA, you learned a new way to use 2FA, or it didn't bother you. [**Open-text question**]*

[**Q7: General feedback**] Lastly, here is some space to leave us feedback or questions, if you like. [**Open-text question**]

Number of participants	308
Gender	
Male	179 (58.1%)
Female	128 (41.6%)
No answer	1 (0.3%)
Age group	
18-19	2 (0.7%)
20-29	54 (17.6%)
30-39	62 (20.3%)
40-49	64 (20.9%)
50-59	69 (22.5%)
60-69	38 (12.4%)
70-79	16 (5.2%)
80-89	1 (0.3%)
No answer	2 (0.01%)
Language	
English	286 (92.9%)
Other	22 (7.1%)
Nationality	
United States	291 (94.5%)
Other	17 (5.5%)

Table C.1: Demographics by Prolific of our participants.

C.1.2 Results

C.1.2.1 Recruiting and Demographics

We recruited our participants via the *Prolific*[1] platform. Prolific collects basic demographic information[2] about their participant pool, to which we added a pre-screening question to select only participants that stated that they use 2FA on at least two different websites. We used Prolific to create a pre-screened participant pool whose demographics are representative of the US population and that has an approval rate of more than 90% on Prolific. From this pre-screened participant pool, we selected a random sample of 30 participants for 3 rounds of pre-testing the survey to remove any ambiguity as much as possible. After pre-testing, we opened the survey for 300 participants from this pool, where 309 in the end successfully finished the survey and we also paid the 9 participants that submitted their completion code too late. For the basic survey (all questions except Q5), which we measured in pre-testing to take about 5 min, we paid £1.20. This corresponds to an hourly wage of £14.40 (≈$16.50/hour). For each additional website in Q5, we estimated 2 min work time and paid £ 0.50 (i.e., £15.00 hourly wage).

One participant in our sample disagreed with our definition of 2FA (Q1), and the results were excluded from the final data set. The demographics as collected by Prolific for the remaining 308 participants are summarized in Table C.1. The average age is 45.2±1.7 years. Most participants are US citizens (94.5%) and have English as a first language (92.9%). Of all participants, 58.1% identified as male, which unfortunately

[1] https://www.prolific.co/

[2] https://researcher-help.prolific.co/hc/en-gb/articles/360009221093-How-do-I-use-Prolific-s-demographic-prescreening-

Question	Q5a	Q5c	Q5d	Q6a	Q6b	**Mean**
Cohen's Kappa	0.781	0.734	0.784	0.835	0.793	0.785

Table C.2: Cohen's Kappa for inter-rater reliability for coding the open-text answers to Q5 and Q6.

differs from the US census 2021 estimation[3] where 50.5% identified as female.

Ethical considerations: For this survey we used the existing infrastructure by a fellow research group at our institution for a series of user studies on Prolific. Since our survey is more restricted in the collection of private data than the remaining user studies, the existing approval by our institutional ERB extended to our survey.

C.1.2.2 Coding Open-Text Answers

Our participants answered open-ended text questions about their 2FA user experiences on different websites. Such qualitative data allows to capture individual perceptions, thoughts, and concerns of users. We used inductive coding (see [123, 80, 208, 134]) to analyze their answers. Two researchers jointly read a randomly sampled subset (25%) of the open-text answers of our participants to Q5 and Q6, discussed them and developed an initial coding scheme for each question where each question was assigned one or more codes. In the next step, the initial codes for each question were merged by axial coding to more abstract codes and the final code books for each question. After this step, two researchers independently assigned a code from the final code books to each answer for Q5 and Q6. We calculated Cohen's Kappa for the finally assigned codes to the answers of Q5 and Q6 to measure the inter-rater reliability (correspondence between the coders). The two coders achieved a satisfactory to near-perfect mean inter-rater reliability, see Table C.2.

C.1.2.3 Qualitative and Quantitative Results

Our participants named, on average 5.06±0.38 websites on which they currently use 2FA (question **Q2a**) and 0.61±0.12 websites on which they stopped using/setting up 2FA (**Q2b**). However, these numbers have to be taken with a grain of salt since a few participants remarked that they could not recall all websites on which they use 2FA or that they could have listed more than the maximum of 20 we asked for. Of all 308 participants, 150 participants (48.7%) indicated a difference in the 2FA experience on at least one of the websites they listed (**Q3**), 158 did not note a difference. Those 150 participants noticed, on average, differences on 1.51±0.31 websites of the previously named ones. For the websites where the participants currently use 2FA, our participants noted differences on average on 1.25±0.32 sites. For websites where they abandoned 2FA, they noticed differences on average on 0.26±0.07 sites.

Of those 150 eligible participants, 112 opted-in (**Q4**) to answer the additional questions in **Q5**. Figures C.1 to C.8 depict the final codes for questions **Q5a** to **Q5d**,

[3] `https://www.census.gov/quickfacts/fact/table/US/PST045221`

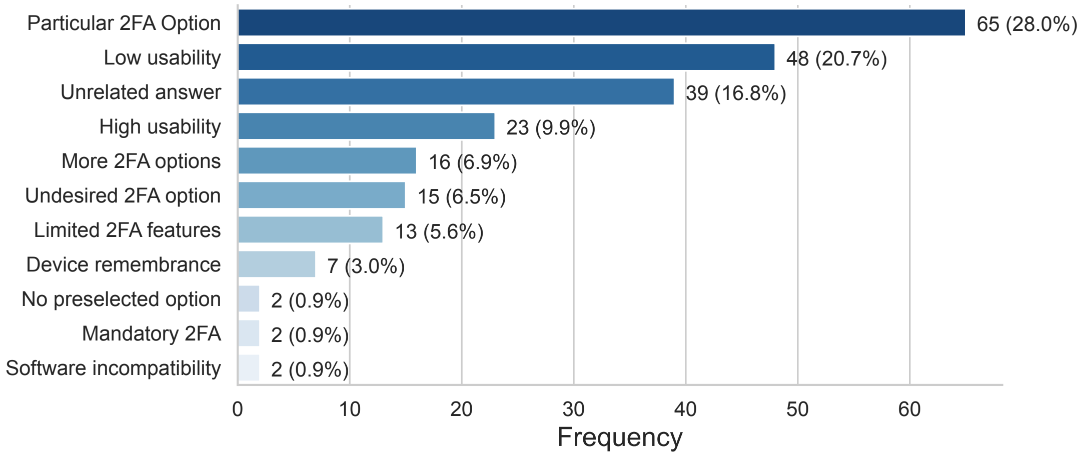


Figure C.1: Codes for **Q5a** ordered by their frequency.

ordered by their frequency, which we elaborate on in the following.

Answers to Q5a: When asked about the differences that they noticed on particular websites (**Q5a**), about a third (28%) of the answers mentioned that the website offered a particular 2FA option (e.g., authenticator app, security keys, or confirmation prompts). For example:

> *"it requires you to get a code from your email, I'm not sure if this is majorly different from other 2FA's but I have not seen it done this way before."* (P4, particular 2FA option)

> *"This is the only site that is tied to Google Authenticator. The other sites that I used, use simple text messages as the secondary method."* (P143, particular 2FA option)

> *"Instead of sending you a verification code it sends you a push notification you have to click "yes it's me" or "decline". Others are always codes."* (P159, particular 2FA option)

Almost a fifth of the answers (20.7%) mentioned that this website differed by having comparatively lower usability of the 2FA experience, such as complex setups with too many steps, worse instructions than other sites, complex settings menu structure, or intransparent policies when 2FA is required during logins. As examples:

> *"Veve sent you an email with a code instead of a text, and this made it harder to get the code. Plus, they had a short time limit that you had to get the code and put it in (Less than a minute or maybe even 30 seconds) and that was hard to make it sometimes."* (P208, low usability)

In contrast, only 9.9% of the answers mentioned that the website sticks out positively by providing higher usability of 2FA, for instance:

> *"Pay Pal explained how to use the 2 step system in more detail and in simpler more understandable terms than anyone else."* (P60, high usability)

> *"Google's 2FA is easy to set up and is very easy when using a hardware key because it supports U2F."* (P304, high usability)

A few answers mentioned that the website differed in that it offered more 2FA options than other sites (6.9%) or that it limits the 2FA features (i.e., 2FA and recovery options) too much (5.6%), sometimes even to the extent that the participants desires another preferred option (6.5%).

> *"They give me many options for how to verify myself."* (P21, more 2FA options)

> *"Okta provides a multide (multiple) of options to enable 2FA on their site. Whenever I have an Okta prompt, I chose to setup an okta verify app that gives me a popup notification to click allow. They also offer text messaging with a code, emails, rotating numbers, and a few more that I don't recall right now as I only use the push notification to their app."* (P50, more 2FA options)

> *"Facebook does not have an SMS part of their 2FA. I prefer getting a quick message to my phone to authenticate my identity."* (P2, limited 2FA features)

> *"I mention sofi because they do not have any backup for 2FA if I lose my phone. Most of the others sites have one-time-use codes we can print out or something like that. I set up two devices with google auth so that's my backup."* (P215, limited 2FA features)

> *"All the other sites let me use an app to provide a code, but ally only allows a text based code. I don't feel it's nearly as secure."* (P108, undesired 2FA options)

> *"I have to download a secondary app, Authenticator that has the 2FA code. I'm generally fine with the quick SMS message, but I don't like to add additional apps to my phone when I don't think it is necessary."* (P148, undesired option)

Lastly, a small number of answers mentioned very concrete technical differences regarding device remembrance support (7; 3.0%), lacking support for selecting the 2FA during login (2; 0.9%), mandating 2FA (2; 0.9%), or incompatibilities with 3rd party software/services due to 2FA (2; 0.9%). Examples for those differences are:

> *"Most other websites seem to allow you to remember a device so you do not need to go through the 2FA process on a trusted device. For some reason Fidelity does not offer that and makes me go through 2FA every time I log in."* (P255, device remembrance)

> *"Vk requires 2FA authentication to log in and doesn't give the option to not use it. A random number calls and the last 4 digits of that number are the authentication code. Sometimes the call takes up to a minute to appear, and you have to remember the number as it calls and type it in quickly."* (P302, mandatory 2FA)

> *"53.com, the website for Fifth Third Bank, recently made 2FA mandatory. Despite my distaste in general for mandatory things, they made the process almost seamless. They seem to have planned very effectively for the change."* (P72, mandatory 2FA)

> *"PAI offers several options for 2FA and offers them every time you log in. All other sites just offer you the choice once and every time after that you have to use the same option."* (P135, no preselected option)

> *"They support FIDO2 which makes login a breeze. Their login is also simple with all your options presented to you to choose, except when FIDO2 is available to make login easier."* (P304, no preselected option)

> *"The interface is very clumsy. The 2FA here also interferes with 3rd party apps like acorn."* (P62, software incompatibility)

> *"Etrade requires an old-school RSA hardware key or special app, plus 2FA makes it hard to connect other services to it, so I turned it off. Their site is old and doesn't support the modern stuff."* (P215, software incompatibility)

A little bit less than a fifth (16.8%) of the answers were unrelated to the question, potentially because the participants misunderstood this or the previous question (**Q4**), e.g., they talked about general security benefits of 2FA or mentioned that the website does not differ from other websites (though we asked in **Q4** which websites *differ* in their opinion). For instance:

> *"First Citizen bank required 1 FA in person while online it required 2 FA via laptop."* (P191, unrelated answer)

> *"enables security by providing 2FA but pretty much similar to other existing 2FA based apps or websites."* (P293, unrelated answer)

Answers to **Q5b**: Figure C.2 summarizes the quantitative data about how participants perceived those differences from **Q5a** in comparison to other websites. Almost have of the answers (47.6%) indicated that those differences made it harder to use the 2FA on that specific website, and only a third of the answers (33.2%) indicated that the website's 2FA was easier to use.

Figures C.3 and C.4 provide a different view of this data. Figure C.3 presents the contingency table between the codes for **Q5a** and the answers for **Q5b**. We measured the statistical association the two variable with Cramer's V (with Bergsma bias-correction),

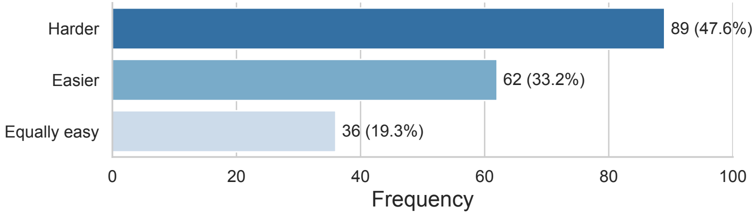


Figure C.2: Quantitative answers for **Q5b** ordered by their frequency.

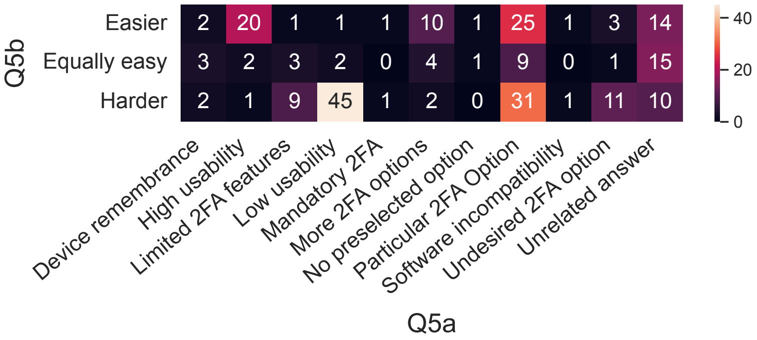


Figure C.3: Contingency table between codes for **Q5a** and answers for **Q5b**.

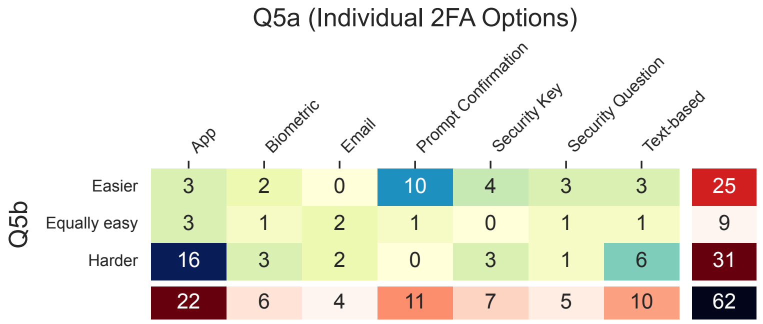
and $V = 0.43$ indicates that there is moderate to a strong association. Not surprisingly, high usability makes the 2FA usage easier and low usability makes it more challenging. However, the data also indicates that offering more 2FA options makes the 2FA usage easier while limiting 2FA options and/or forcing users to apply an undesired option makes the website harder to use. Figure C.4 splits the “particular 2FA option” code into the options identified in the participants’ answers and shows the contingency table between individual options and the answers to **Q5b**. The figure indicates that app-based two factor authentication is in particular correlated with a harder 2FA usage by our participants, while confirmation prompts are perceived as an easier to use 2FA option.

Answers to **Q5c**: Figure C.5 shows the final codes for the answers to question **Q5c**, ordered by frequency. When asked how participants behaved in reaction to the differences in 2FA on a particular website, almost a third (28.9%) responded that they did not change their behavior. For example:

> *“I log into this website out the same as I normally would, I just find the process much more convenient than other options.”* (P270, unchanged usage behavior)

> *“I’ve had no change in usage. But I do find it much easier to use 2FA instead of passwords.”* (P220, unchanged usage behavior)

> *“No change to behaviour, but I do like it more.”* (P305, unchanged usage behavior)



Q5a (Individual 2FA Options)

Q5b	App	Biometric	Email	Prompt Confirmation	Security Key	Security Question	Text-based	
Easier	3	2	0	10	4	3	3	25
Equally easy	3	1	2	1	0	1	1	9
Harder	16	3	2	0	3	1	6	31
	22	6	4	11	7	5	10	62

Figure C.4: Contingency table between individual 2FA options mentioned in answers for **Q5a** with code "Particular 2FA option" and answers for **Q5b**.

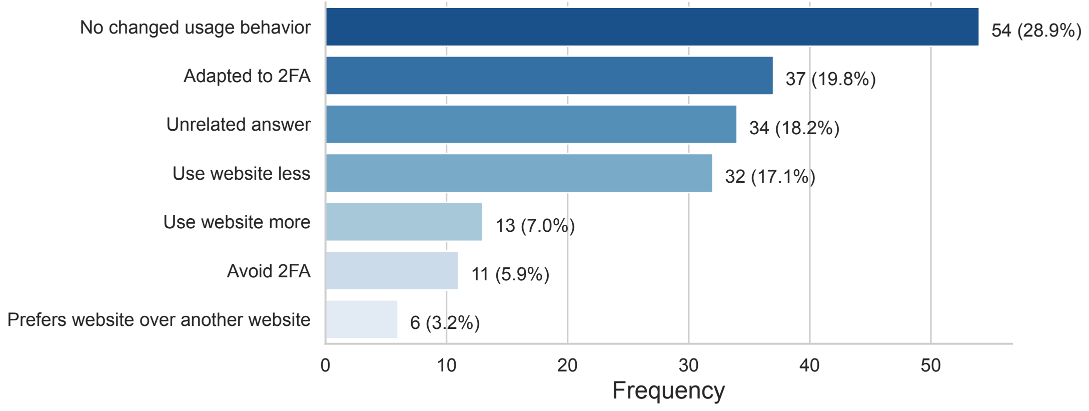


Figure C.5: Codes for **Q5c** ordered by their frequency.

Among the answers that described a changed usage behavior, most commonly, the participants indicated to have adapted to the particular 2FA of the website (19.8% of all answers), e.g., learning how to use a prior unfamiliar 2FA option, or to use the website less frequently (17.1%), in some cases even to the extent of abandoning the website or switching to a partner app for the service.

> *"I went through all of the learning steps and I watched video tutorials."* (P34, adapted to 2FA)
>
> *"I log in as much as I need to but prefer that they would let me pick the authenticator app or at least have more options available but since it has a time on it I try to get on and finish what I'm doing quickly without leaving my computer."* (P296, adapted to 2FA)
>
> *"I use the website on my computer less and use the mobile app more often."* (P255, use website less)
>
> *"I log on to the website less frequently than I typically would because the security questions get annoying after time."* (P296, use website less)

A few participants described that they developed strategies to avoid having to use 2FA (5.9%), e.g., trying to keep the logged in a session open as long as possible.

> *"I just try to remain logged in on Coinbase so I don't have to go through it again."* (P97, avoid 2FA)

> *"I try to not log out."* (P48, avoid 2FA)

> *"I try very hard not to have to like full login as in just leave my devices logged INTO google at ALL times so I can avoid the hassle of having all my devices go HOLD UP even for a moment."* (P201, avoid 2FA)

Some participants with positive experiences with the 2FA on this particular site also described that they now use the website more often (7.0%) or, prefer this website over another service (3.2%). For example:

> *"I log into this website more since it is an easy 2FA."* (P230, use website more often)

> *"I login in more. It's quick and simple."* (P285, use website more often)

> *"I login more because it is so quick and I just have to plugin or tap my yubikey."* (P304, use website more often)

> *"I always choose it over the Barclays site."* (P167, prefer website over another)

> *"I log into Chase more and bank with them because its easier rather than use a lot of the other banks. I like ease of use and I am a simple person."* (P299, prefer website over another)

Lastly, about a fifth of the answers (18.2%) were unrelated to the question, e.g., they referred to a companion app instead of the website, reported about an unrelated technical problem (e.g., a website was down), or related to problems with general security/privacy policies and risk-based authentication. For example:

> *"I couldn't use Outlook outside of work for awhile."* (P37, unrelated answer)

> *"I think I've started to think more about the data I share with other sites that feel less secure."* (P95, unrelated answer)

Figure C.6 relates the differences pointed out in **Q5a** with answers that mention that the participant had to adapt to 2FA on the specific website, tries to avoid the 2FA when possible or use the website less. The data indicates that low usability, in contrast to other websites, caused the participant to use the website in half of the cases less. Limited 2FA features and undesired 2FA options almost equally caused participants to either adapt or to use the website less.

Figure C.7 depicts the contingency table for code *Particular 2FA option* from Figure C.6 broken down into individual 2FA options mentioned in answers to **Q5a**.

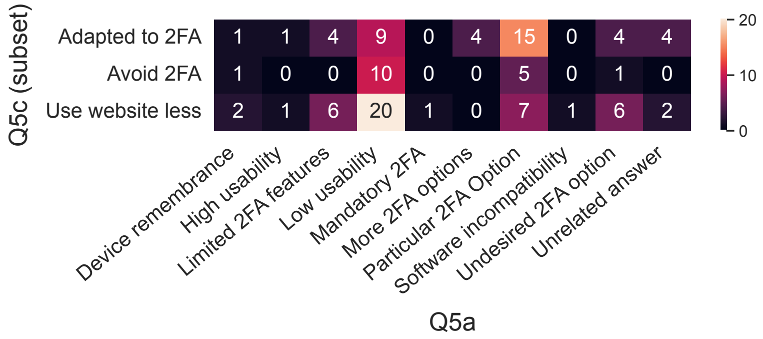


Figure C.6: Contingency table between codes for **Q5a** and codes *Adapted to 2FA*, *Avoid 2FA*, and *Use website less* for **Q5c**.

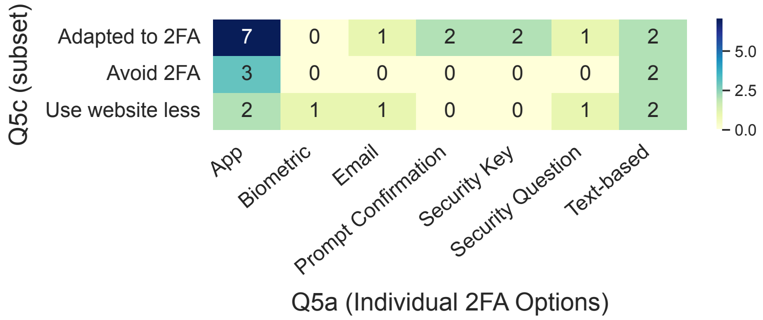


Figure C.7: Contingency table between individual 2FA options mentioned in answers for **Q5a** with code "Particular 2FA option" and codes *Adapted to 2FA*, *Avoid 2FA*, and *Use website less* for **Q5c**.

App-based 2FA has been mentioned most often in combination with *Adapted to 2FA*, *Avoid 2FA*, and *Use website less*, where the majority of participants adapted to this option. Only for app-based 2FA and text-based 2FA, our participants mentioned that they developed strategies to avoid two-factor authentication when possible. Among our participants, all 2FA options except prompt confirmation and security keys have caused at least one participant to use a website less.

Answers to **Q5d**: Figure C.8 shows the codes for answers to question **Q5d**, ordered by frequency. When asked what participants would recommend other websites to adopt or avoid from 2FA on the website **Q5** is currently asking about, a fifth of the answers (19.8%) suggested a general way to avoid the 2FA of that website. This referred to general usability problems of 2FA, e.g., in the settings, setup, or usage. For example:

> *"Avoid creating too many extra steps. Others like Yahoo do this seamlessly."* (P7, avoid 2FA of this website)

> *"Itunes should make their process simpler and and not force a re-login across different platforms like Google does, which is easier."* (P279, avoid 2FA of this website)

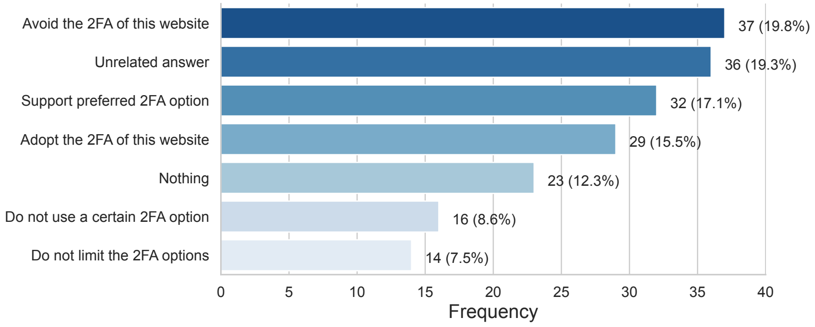


Figure C.8: Codes for **Q5d** ordered by their frequency.

> *"I do not think other websites should use the 2FA in a similar way to Udemy."* (P296, avoid 2FA of this website)

> *"yes, use standard 2fa, don't try and reinvent the wheel."* (P43, avoid 2FA of this website)

Fewer answers noted very particular things that other websites should avoid, such as a certain 2FA option (8.6%) or limiting the 2FA options (7.5%).

> *"Most other sites give multiple, more secure options for 2FA. Ally feels outdated and behind limiting you to either a text code or email code. Some sites allow multiple types of 2FA enabled at the same time (app, email, etc.) but Ally does not."* (P108, do not limit the 2FA options)

> *"It's simple: Give users multiple different choices for 2FA!"* (P290, do not limit the 2FA options)

> *"No other websites should use the text message code to authenticate an account."* (P270, do not use a certain 2FA option)

> *"avoid SMS text messages."*(P308, do not use a certain 2FA option)

Almost a fifth of the answers (17.1%) concretely recommend that other websites should adopt their preferred 2FA option.

> *"I think all websites should have a click button or a push notification (whether app or to text) that is a one click verification."* (P50, support preferred 2FA option)

> *"I think more sites should offer QR code login."* (P244, support preferred 2FA option)

Of all answers, 15.5% made general recommendations where other websites should adopt something from the two-factor authentication experience of this particular website. For example:

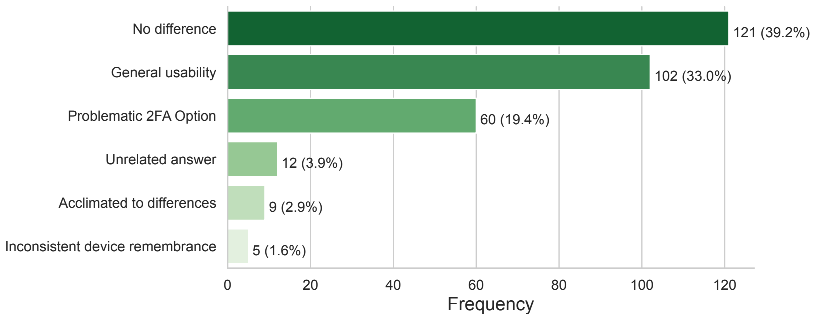


Figure C.9: Codes for **Q6a** ordered by their frequency.

> *"All sites should seek out an easier way like Yahoo where you only have to click OK rather than re-enter numbers."* (P7, adopt the 2FA of this website)

> *"They should adopt the 2FA standards that Google support because it makes it much more easier."* (P304, adopt the 2FA of this website)

Only 12.3% of the answers did not make a recommendation about what to avoid or adopt on other websites. About a fifth (19.3%) of the answers were unrelated to the question, since they recommend general features unrelated to 2FA or made general statements about 2FA:

> *"I think, overall, that most sites dealing with personal/financial/health data should all be 2FA."* (P125, unrelated answer)

> *"They should allow for a personally chosen user ID."* (309, unrelated answer)

Answers to **Q6a**: In question **Q6a**, we asked every participant if they recall a specific situation in which 2FA was problematic due to differences in the 2FA user experience to what they were used to. Figure C.9 lists the codes for the answers to **Q6a**.

About 40% of the participants did not recall such a situation or indicated that they did not notice problematic differences.

> *"I cannot think of any situation where it was difficult, I have always found it pretty easy."* (P30, no difference)

> *"No, they have all seemed about the same."* (257, no difference)

And 3.9% of the answers were unrelated to the question **Q6a**, e.g., referring to general problems in creating an account or logging in with passwords.

> *"I forgot one of the passwords, (I used a unique one only for this). I misentered something I guess."* (P42, unrelated answer)

Of all 308 participants, the remaining 176 participants (57.1%) recalled a problematic situation. About half of them (102; 33% of all answers) mentioned that a website had general usability problems in their 2FA experience, but the answer did not specify if and how this differs from the participants' usual 2FA experiences.

> *"My only thought is Paypal and that it isn't consistent. I probably need to check the settings. I have not had to do that with any of the other sites."* (P57, general usability)

> *"Yes, see previous. Citi changed their method without informing me."* (P71, general usability)

About a fifth of all answers (19.4%) mentioned a problematic 2FA option in contrast to other websites, such as a custom 2FA option by that service, an unfamiliar 2FA option, or being required to use an undesired 2FA option. For example:

> *"I usually don't continue a 2FA set-up if the 2FA involves a text. I prefer to use the authenticator app or my email."* (P164, problematic 2FA option)

> *"Not really. Most use txt messages, one uses email. Amazon also uses texts. The Amazon one made you click on the link in the text, which I don't like."* (P131, problematic 2FA option)

> *"I find it to be very annoying when websites have me download additional software to my phone such as an Authenticator app in order to go through the 2FA process. I can't stand downloading additional apps when a simple text or email will do."* (P200, problematic 2FA option)

> *"In the past, Verizon website had multiple additional steps of 2fa. In addition to 2fa itself, one also needed to remember a picture and code word combination that was set up earlier. I found both difficult to remember and was happy when Verizon switched to 2fa without any additional steps."* (P83, problematic 2FA option)

Five participants (1.6%) recalled a situation in which the device remembrance logic of a website worked differently than they expected it or was used to it.

> *"2FA was more annoying on Runescape because you weren't able to select an option for your device to be remembered. So, each time I would log in I'd have to grab my phone, open the authentication app, and input the number sequence."* (P82, inconsistent device remembrance)

> *"Some only use 2FA once unless it is a new device, Others use it everytime."* (P294, inconsistent device remembrance)

Nine participants recalled a situation or even general inconsistencies in 2FA but directly stated that they got used to it and do not see a problem (anymore).

> *"I'm used to the differences; they seem kind of random: sometimes a texted code, sometimes a fingerprint; sometimes my use of LastPass Authenticater. All work well enough."* (P36, acclimated to differences)

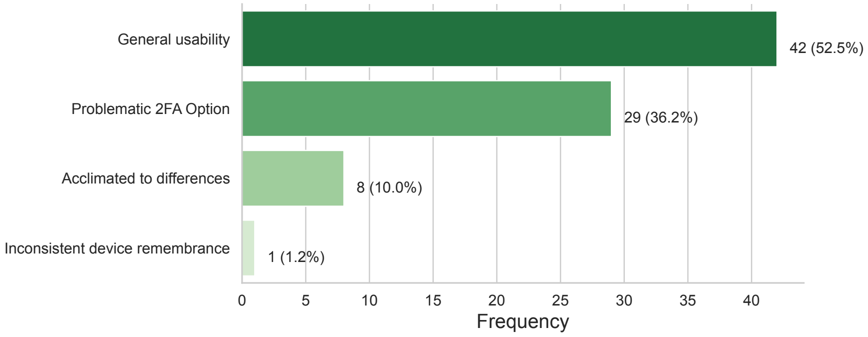


Figure C.10: Codes for **Q6a** only for participants that did not selected websites in **Q3**.

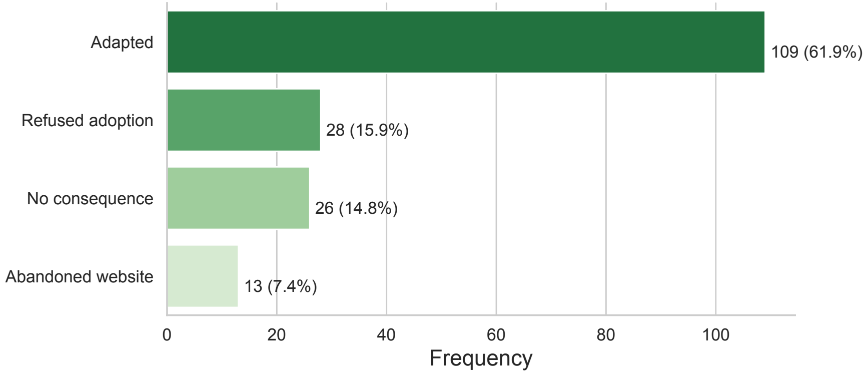


Figure C.11: Codes for **Q6b** ordered by their frequency.

> *"Probably a 2FA experience I had with work. We use our own authenticator app because I work for a bank, and it is a little complex. For example you have to log into your computer, put in your password, then go to an app on your phone, enter the password again, it will throw up a generated code which you only have 20 seconds to type in. It took some getting used to for sure."* (P290, acclimated to differences)

Of those 176 participants that recalled a situation, 80 did not previously in **Q3** indicate differences in their 2FA experiences between websites (i.e., they did not go through **Q4** and **Q5**). Figure C.10 summarizes their answers to **Q6a** separately. Their answers show that although they did not initially indicated differences between 2FA experiences, they recalled in many cases, a situation in which a website had 2FA usability issues or a problematic 2FA option.

Answers to **Q6b**: Figure C.11 summarizes the answers for question **Q6b** from those 176 participants that recalled a problematic situation in **Q6a**. The majority (61.9%) of those participants explained that they had to adapt to the situation, e.g., by switching to a less preferred 2FA option, fighting their way through despite the problems, or learning how to use a new 2FA option.

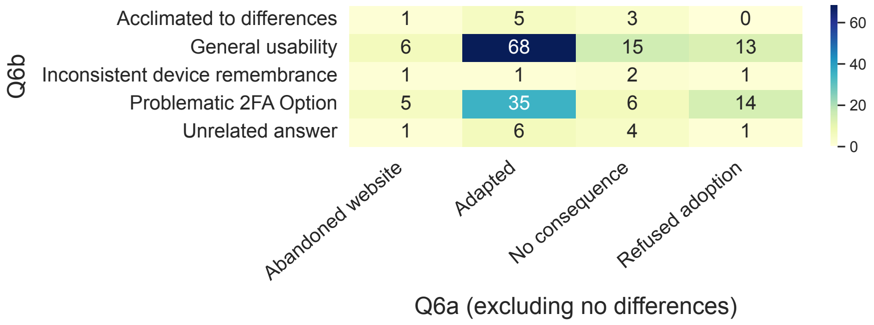


Figure C.12: Contingency table between codes for **Q6a** (excluding *No differences*) and **Q6b**.

> *"I generally like to opt-in for SMS messaging for 2FA, so if that is not offered, I will adjust."* (P193, adapted)

> *"I still used 2FA because it was required, but I would have disabled it if I could. I also log into it a little less since I don't want to deal with getting the code."* (P208, adapted)

However, several participants responded that they refused the adoption of 2FA in this situation (15.9%) or abandoned the website due to the indicated problems (7.4%).

> *"I chose not to sign up, and didn't use the site."* (P26, abandoned website)

> *"I stopped using the site. If I can't trust their authentification info, I can't trust the site."* (P136, abandoned website)

> *"Stopped banking with HSBC."* (P262, abandoned website)

> *"I had to email them then once I got back on the account I shut 2FA off."* (P210, refused adoption of 2FA)

> *"If I couldn't use VOIP I removed some 2FA."* (P69, refused adoption of 2FA)

> *"I aborted setting up 2FA on those sites."* (P242, refused adoption of 2FA)

Figure C.12 depicts the contingency table between codes for **Q6a** (excluding *No differences*) and **Q6b**. While most participants adapted to problematic 2FA options or general usability issues, 14 and 13 participants, respectively, refused to adopt 2FA. One participant refused adoption of 2FA because of friction with the device remembrance policy that differed from other websites.

Of all 176 participants, 26 (14.8%) answered that there were no consequences for them.

C.1.3 Conclusion of our survey

In regards to the question of whether users had negative experiences transferring their 2FA experiences between websites and whether this has stopped them from enabling or continuing to use 2FA, 28 (9.1%) of our 308 participants mentioned for at least one website, which differed in their opinion from others in its 2FA experience (**Q5**), that they use this website less due to this 2FA experience. Additionally, 41 (15.9%) of the participants recalled a concrete situation with 2FA that was challenging because the 2FA experience differed from what they were used to (**Q6**) and, as a result, they abandoned the website or refused adoption of (a specific) 2FA option. Taken together, 60 (19.5%) of our participants reported using a website less, abandoning a website, or refusing adoption of (a specific) 2FA option. Of those, 28 (9.1% of all participants) refused adoption due to differences in the usability of 2FA in contrast to other websites, undesired/unfamiliar/custom 2FA options, or in one case due to an inconsistent device remembrance policy.

C.2 Website Details

Table C.3 provides additional details about the websites in our data set, including their Tranco rank, rank category (see Section 4.7.2.3), naming and location of 2FA settings, description of 2FA, and form of device remembrance.

C.3 Details on Pairwise Hamming Distances

As explained in Section 4.7.2.1, we compare the 85 websites in our dataset using pairwise Hamming distance between the 14 non-conditional factors of each website. Figure C.13a depicts the (cumulative) frequency distribution of the pairwise Hamming distances between all websites in our data set, where a distance of 0 means equality in all factors and a distance of 1 means complete inequality of all 14 factors. This distribution is very symmetrically (skew=0.062) and only slightly heavy-tailed (kurtosis=-0.112), but an omnibus test of normality [171] ($p > .05$) indicates that it is not Gaussian. Further, Figure C.13b shows each website's frequency distribution for minimum, mean, and maximum distance. The average website in our data has a minimum Hamming distance of 0.16 ± 0.02 (for a confidence interval of 95%), a mean distance of 0.46 ± 0.01, and a maximum distance of 0.75 ± 0.01. In other words, the average website in our data set differs on average in 6–7 out of 14 factors from the other websites and differs on average in at least 2–3 factors. Nevertheless, no pair of websites has a distance larger than 0.86, i.e., there are always two factors identical for each pair of websites.

C.4 High-level View of Clusters

Figure 4.13 provides a less noisy view of the clusters depicted in Table 4.1 to see the clusters' structure easily. Similarly, Figure 4.14a depicts only the *non-conditional* factors for which we identified six *inter-clusters*, and Figure 4.14b depicts only the *conditional* factors for which we found three *subclusters* or *intra-clusters*.

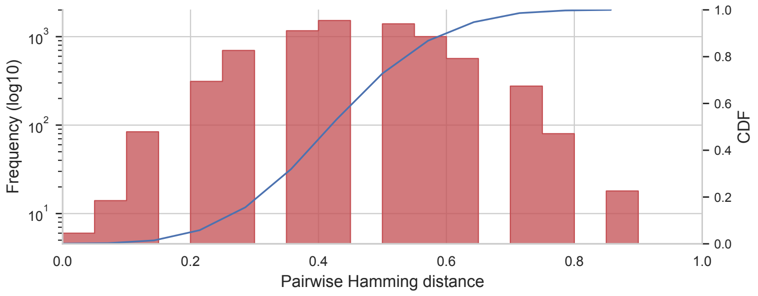
Category	**Website**	**Rank**†	**Rank Cat.**	**Naming**	**Location**	**Descr.**	**Remembr.**
Backup and Sync	backblaze.com	6973	Long tail	other	Security / Account	—	Trust
	dropbox.com	56	Top-500	2SV	Security / Account	Security	Trust
	evernote.com	412	Top-500	2SV	Security / Account	Security	Remember
	icloud.com	111	Top-500	2FA	Security / Account	—	Trust
	jottacloud.com	149805	Long tail	2FA	Security / Account	Security	—
	mega.io	21990	Long tail	2FA	Security / Account	Security	—
	synology.com	3267	Top-4000	2SA	Security / Account	Security	Remember
Cloud Computing	digitalocean.com	1573	Top-4000	2FA	Security / Account	Security	—
	laravel.com	3424	Top-4000	2FA	Security / Account	—	—
Communi-cation	basecamp.com	1483	Top-4000	2FA	Security / Account	Security	—
	discord.com	460	Top-500	2FA	Security / Account	Security	—
	mailchimp.com	222	Top-500	2FA	Security / Account	Security	Skip
	zoom.us	34	Top-500	2FA	Security / Account	—	—
Crowdfunding	kickstarter.com	292	Top-500	2FA	Security / Account	Security	Remember
Crypto-currencies	binance.com	466	Top-500	2FA	Security / Account	Security	—
	bitfinex.com	2870	Top-4000	2FA	Security / Account	Security	—
	blockchain.info	8244	Long tail	2SV	Security / Account	Device	—
	bybit.com	9189	Long tail	2FA	Security / Account	—	—
	kraken.com	3950	Top-4000	2FA	Security / Account	Security	—
Developer	atlassian.com	671	Top-4000	2SV	Security / Account	Device	—
	github.com	35	Top-500	2FA	Security / Account	Security	—
	gitlab.com	712	Top-4000	2FA	Security / Account	Security	—
	unity.com	2480	Top-4000	2FA	Security / Account	—	—
Domains	easydns.com	36665	Long tail	2FA	Security / Account	—	—
	gandi.net	3644	Top-4000	2FA	Security / Account	Security	—
	hover.com	5585	Long tail	other	Security / Account	Security	—
	namecheap.com	852	Top-4000	2FA	Security / Account	Security	—
Email	google.com	1	Top-500	2SV	Security / Account	Security	Remember
	yahoo.com	18	Top-500	2SV	Security / Account	Security	Remember
	zoho.com	365	Top-500	MFA	Security / Account	Security	Skip
Entertainment	twitch.tv	76	Top-500	2FA	Security / Account	Security	Remember
Finance	xero.com	1238	Top-4000	MFA	Security / Account	Security	Skip
	youneedabudget.com	8478	Long tail	2SV	Security / Account	Security	—
Gaming	blizzard.com	2258	Top-4000	other	Security / Account	Device	—
	ea.com	789	Top-4000	other	Security / Account	Security	—
	epicgames.com	920	Top-4000	2FA	Security / Account	Security	Remember
	playstation.com	707	Top-4000	2SV	Security / Account	Security	—
	roblox.com	248	Top-500	2SV	Security / Account	Security	Trust
Government	va.gov	1102	Top-4000	MFA	Security / Account	Security	—
Health	23andme.com	6345	Long tail	2SV	Security / Account	—	—
	runsignup.com	7326	Long tail	MFA	Other	Security	—
Hotels/Acom.	booking.com	139	Top-500	2FA	Security / Account	—	—

continued on next page

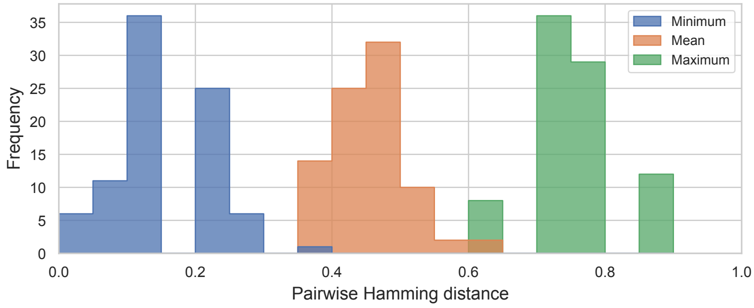
Category	Website	Rank†	Rank Cat.	Naming	Location	Descr.	Remembr.
Identity Management	1password.com	3012	Top-4000	2FA	Other	Security	—
	bitwarden.com	10002	Long tail	other	Security / Account	Security	Remember
	id.me	6012	Top-4000	other	Security / Account	Security	—
	keepersecurity.com	12065	Long tail	2FA	Security / Account	—	Skip
	lastpass.com	1333	Top-4000	other	Security / Account	Security	Trust
	orcid.org	1697	Top-4000	2FA	Security / Account	Security	—
	roboform.com	5312	Long tail	other	Security / Account	Security	—
IoT	arlo.com	10953	Long tail	2SV	Other	Security	—
	ifttt.com	2847	Top-4000	2SV	Security / Account	Security	—
Legal	clio.com	20552	Long tail	other	Security / Account	Security	—
	docusign.com	815	Top-4000	2SV	Security / Account	Security	Remember
Other	adobe.com	23	Top-500	2SV	Security / Account	Security	Other
	opera.com	133	Top-500	2FA	Other	Security	—
Payments	paypal.com	75	Top-500	2SV	Security / Account	Security	Trust
	stripe.com	630	Top-4000	2SA	Other	Security	—
Remote Access	join.me	12578	Long tail	2SV	Security / Account	Security	—
	logmein.com	2714	Top-4000	2SV	Security / Account	Security	Trust
	realvnc.com	10442	Long tail	2SV	Security / Account	Security	—
	teamviewer.com	455	Top-500	2FA	Other	Security	—
Retail	amazon.com	17	Top-500	2SV	Security / Account	Security	Remember
	ebay.com	70	Top-500	2SV	Security / Account	Security	—
	etsy.com	107	Top-500	2FA	Security / Account	Device	Other
	newegg.com	1141	Top-4000	2SV	Security / Account	Security	Remember
Security	bitdefender.com	2198	Top-4000	2FA	Security / Account	Security	Trust
	cloudflare.com	79	Top-500	2FA	Security / Account	Security	—
	digicert.com	158	Top-500	2FA	Security / Account	Security	Remember
	norton.com	872	Top-4000	2FA	Security / Account	Security	Trust
	virustotal.com	2295	Top-4000	2FA	Security / Account	Security	Remember
Social	facebook.com	4	Top-500	2FA	Security / Account	Device	—
	instagram.com	7	Top-500	2FA	Security / Account	Device	—
	linkedin.com	9	Top-500	2SV	Security / Account	Security	Remember
	reddit.com	32	Top-500	2FA	Security / Account	Security	—
	tumblr.com	45	Top-500	2FA	Security / Account	Security	—
	twitter.com	6	Top-500	2FA	Security / Account	Security	—
	vk.com	50	Top-500	2SV	Security / Account	Security	Remember
Task Management	airtable.com	2052	Top-4000	2SA	Security / Account	Security	—
	clickup.com	7723	Long tail	2FA	Security / Account	Security	—
	clubhouse.io	23778	Long tail	2FA	Security / Account	Security	—
	meistertask.com	18630	Long tail	2FA	Security / Account	Security	—
	toodledo.com	33344	Long tail	other	Security / Account	Security	—
Utilities	callcentric.com	75387	Long tail	other	Other	Security	Remember
	coned.com	22715	Long tail	2SV	Security / Account	Security	—
VPN	airvpn.org	64852	Long tail	other	Security / Account	Security	—

— = not applicable; † based on Tranco [125, 184]

Table C.3: Details about websites in our data set.



(a) Overall distribution of distances.



(b) Min, max, and median distances.

Figure C.13: Frequency distributions of pairwise Hamming distances of non-conditional comparison factors between all websites in our dataset.

C.5 Opinionated Separation of Comparison Factors

We conducted an expert assessment of our comparison factors with local experts to create an opinionated separation of our factors based on their impact on *user experience*, *security*, *both UX and security*, or *neither*. In addition to the authors, we invited four usable security researchers from two other research groups at our institution to participate as experts in this assessment. The invited experts had extensive experience in usability studies and were familiar with web authentication. We presented each expert with our comparison factors with their definition. In an online survey, each expert separately decided which of the four categories each factor falls and provided a short rationale for their categorization. The authors discussed the individual assessments and assigned the final category to each factor. This decision was based on a clear majority among the experts, evidence from the literature and best practices, rationales, and personal opinion by the authors. Table 4.6 lists the categories for each factor, and in the following Section C.5.1 we briefly summarize the rationales that led to this separation. A point worth noting from our discussion is the definition of *"impact on security."* Improvements in usability and user experience are very often entangled with the users' security—for example, nudging more users to adopt 2FA will increase the overall security; streamlining the setup of 2FA options lowers the friction and can cause a higher adoption; or highlighting certain options, like security keys, might persuade users to adopt stronger options over, e.g., text-based OTP. We decided to limit “security” to the *direct* impact on the security of 2FA and the user’s account, and to assign the “security” or “user experience” categories when either category clearly outweighs the other (in our opinion). Afterward, in Section C.5.2, we present additional views on the factor clusters with separated comparison factors (see also Section 4.7.2.4).

C.5.1 List of Comparison Factors

Promotion [UX]: By making 2FA support more visible, users might perceive the website as more secure. The promotion dialog can be the entry point for the 2FA journey. Golla et al. [82] recently showed that a certain kind of messaging and UX design patterns can effectively improve 2FA adoption. However, the promotion can only convince users to start their 2FA journey, but its impact on a security depends on other factors that determine the success of this journey. Thus, we think the impact on UX outweighs the impact on security.

Non-Optional [Both]: Abbott and Patil [4] have shown that mandatory 2FA can increase users’ frustration, affecting the UX. However, when every user account is additionally secured with 2FA has been shown to be an effective defense against account takeovers [52].

Common-Naming-and-Location [UX]: By providing users with a familiar interface/workflow, the website facilitates the discovery of 2FA and reduces user frustration. Following a common naming and location also fits the usability heuristic “Consistency and standards” [144].

Descriptive-Notification [UX]: Similarly to Promotion, this can help users to make more informed decisions and might nudge them to start their 2FA journey at this point but does not directly affect the security of the user’s 2FA.

Additional-Information [Neither]: While more information can help users better understand 2FA, this particular additional information is not presented within context (like descriptive notification or option-specific information) and requires users to take a detour to find it. Thus, the impact on the user experience is considered only marginal, and it does not directly affect the security of 2FA.

Option-Specific-Information [UX]: This factor implements UX guidelines to provide adequate contextual help [35] and timely guidance [115]. A better understanding of the 2FA options definitely affects user experience. If users have a free choice of 2FA options, this understanding *might* also lead them to select a more secure 2FA choice. However, However, as shown in the literature, we consider the impact on UX to outweigh these potential security benefits since such information alone do not guarantee that users *actually* choose (*exclusively*) stronger 2FA options.

Step-Wise-Instructions [Both]: Similar to option-specific information, this implements UX guidelines to provide adequate contextual help, break down complex tasks, and offer assistive interfaces without the need to break the user’s flow [115, 35, 214, 147] and, hence, improve the UX of the 2FA setup. Additionally, clear instructions should reduce the risk of mistakes and errors that might affect security. Fewer errors should also increase the success rate of users in finishing the 2FA setup.

Multiselection [Both]: Offering users multiple options allows them to choose the one that fits them best, affecting their 2FA experience. This is particularly clear from the answers to our survey (see Appendix C.1). In contrast to option-specific information, multiselection affects security directly since it reflects whether users can actually choose between options with potentially different security levels.

Grouped-Setting [UX]: A single location for the settings decreases the burden on the user. This is also summarized in UX guidelines, such as the "gestalt principles" for the common region and proximity [102, 215]. Not grouping the 2FA settings *might* cause users to miss stronger 2FA options. However, similarly to option-specific information, we consider the already demonstrated impact on UX to outweigh this potential security implication.

No-Enforced-Options [Both]: Prior work [32] has shown that enforcing a certain 2FA option while not communicating that additional 2FA options become available afterward has confused users that were explicitly looking for a specific 2FA option, which differed from the enforced one. We also consider this an illustration of Norman's Gulf of Execution [149]. Thus, this factor can have an impact on the user experience. Further, forcing users to set up a "weak" 2FA option first, e.g., SMS-based OTP, directly affects security by preventing by-design that users can achieve the strongest possible account security.

Selectable-Primary-Option [UX]: Allowing users to personalize their 2FA login experience is in accordance with UX guidelines, increasing the usability of the login process and, hence, reducing friction. Setting a primary option does not affect the choices of 2FA options, just their order during login. Hence, the factor does not affect the 2FA security.

Settings-Changed-Verification [Security]: The verification authorizes a security-critical change in the settings, hence, this factor has an impact on the 2FA security of the website. Assuming that the authorization is done with the default authentication, e.g., password plus any second factors, this authorization does not differ significantly from a regular login. We assume that this is the usual scenario and, hence, the impact on security outweighs the impact on user experience.

Settings-Changed-Notification [Both]: Notification about the successful change of the settings reassures users, and we consider it part of the UX best practice for visibility of the system status [88]. A notification also informs the user, even if they are not interacting with the system currently, about potentially unauthorized actions and allows them to take remediation steps.

Confirm-Successful-Setup [UX]: This factor relates to the UX since confirming a successful setup provides visibility of the system status [88], illustrates Norman's Gulf of Evaluation [149], and also addresses the peak-end rule for UX [215]. It also enhances the experience by reducing the chance that the user accidentally locks themselves out of their account. This feedback could also enhance security if the user notices failures and mistakes easier, e.g., incomplete or failed TOTP setup that leads the user to incorrectly believe that their account is secured. However, we considered the impact on UX to outweigh this potential impact on security.

Informed-2FA-Recovery-Options [UX]: Offering users an (easy) way to reduce the chance of being locked out of their account increases the usability and the experience. The availability of recovery options as a "fail-safe" might encourage users to test 2FA.

Recovery options can affect the account security if they are weaker than the 2FA option and provide an easier attack surface. However, in the end, considering the answers to our survey in Appendix C.1, the impact on UX when users are locked out of their accounts seems to outweigh (for now) the potential risks that a weak recovery option poses.

Enforced-2FA-Recovery-Setup [**UX**]: Mandating a recovery option can help users to reduce the risk of account lock-out but also might increase friction and annoy users. The argument for the impact on security remains the same as for Informed-2FA-Recovery-Options.

Device-Remembrance [**UX**]: Ciolino et al. [32] reported that an unexpected device remembrance policy frustrated their participants. Reynolds et al. [164] as well as our survey results corroborate this issue. On the other hand, device remembrance reduces the frequency of 2FA, which some of our survey participants appreciated and considered an enhancement of their UX. Device remembrance can create an attack surface if the attacker has access to a trusted device (e.g., a private device or an accidentally trusted shared/public computer). However, we consider the impact on UX for all users to outweigh the security drawbacks in this narrow threat model.

No-Preselected-Option [**UX**]: Participants of Ciolino et al. [32] expressed a desire to personalize the 2FA experience and also UX guidelines recommend ways to allow users to customize the system to their preferences. Like Selectable-Primary-Option, offering users to choose between their available 2FA options during login does not affect which "weaker" or "stronger" options of 2FA the user has set up, just their order during login. Hence, the factor does not affect the 2FA security.

Informed-Deactivation [**Both**]: Allowing users to deactivate 2FA is in accordance with UX guidelines, like the heuristic for user control and freedom [166], while not allowing them to deactivate can frustrate users. On the other hand, allowing users to deactivate 2FA also directly allows them to make their accounts less secure.

Deactivation-Verification [**Security**]: Like Settings-Changed-Verification, a security-critical change in the settings has to be authorized. We assume that also here, the authorization takes place via a default authentication (e.g., password plus any configured 2FA option) that does not add additional friction for the user.

Deactivation-Notification [**Both**]: Notification about the deactivation of 2FA can reassure users, and we consider it part of the UX best practice for visibility of the system status [88]. Like Settings-Changed-Notification, an out-of-band notification also informs the user about potentially unauthorized actions and allows them to take remediation steps.

Communicate-Successful-Deactivation [**UX**]: Clearly communicating the deactivation of 2FA follows UX best practice for communicating the system status [88]. While this feedback could warn a user in case they accidentally deactivated 2FA and put their account at higher risk, we consider this rather unlikely and the UX impact to outweigh the potential security benefits.

C.5.2 Pairwise Hamming distance and Factor Clusters

We repeat the pairwise Hamming distance measurement and the factor clustering from Section 4.7.2 with separated sets of comparison factors. We split the factors into those purely UX-related (i.e., [UX]) and those with security-relevance (i.e., [Security] and [Both]) from Appendix C.5.1. We further differentiate conditional versus non-conditional factors. This results in a set of four disjoint sets of comparison factors for clustering: *Non-conditional-UX*, *Non-conditional-Security*, *Conditional-UX* and *Conditional-Security*.

Pairwise Hamming distance: We repeated measuring the pairwise Hamming distances from Appendix C.3 for the sets of separated factors. Considering only *non-conditional-UX* factors, the average website has a minimum Hamming distance of 0.1 ± 0.02, a mean distance of 0.47 ± 0.02, and a maximum distance of 0.86 ± 0.01. For the set of *non-conditional-security* factors, the average website has a minimum distance of 0.04 ± 0.02, a mean distance of 0.44 ± 0.02, and a maximum distance of 0.90 ± 0.01. Thus, compared to all factors, the general impression is that these separate sets of factors do not exhibit a better overall consistency across the websites in our data set.

Factor clusters: Similar to the high-level view of our clusters with all comparison factors in Appendix C.4, we provide different views on clusters based on the separated sets of factors. We again computed the mean Silhouette Coefficient for each set of factors for different clusters with KModes to determine the best number of clusters to describe our data set. For *Non-conditional-UX* comparison factors, we computed 2 and 5 to be the best number of clusters. Since, with 2 clusters, the websites were only clustered by their strategy for *Option-specific-information*, we chose 5 as the more expressive number of clusters. Figure C.14 illustrates the result of the clustering. We computed 10 as the best number of clusters for *Conditional-UX* comparison factors, illustrated in Figure C.15. For *Non-conditional-security* comparison factors, we computed 9 clusters as the best number of clusters. Figure C.16 illustrates the resulting clusters. Moreover, for *Conditional-security* comparison factors, Silhouette testing showed 8 as the best number of clusters, see Figure C.17. We also combined the non-conditional and conditional clusters to get the complete picture of the purely UX-related and security-related clusters. The results are depicted in Figures C.18 and C.19, however, the high number of intra-clusters and, for security-related factors, inter-clusters make a meaningful interpretation of this noisy view hard.

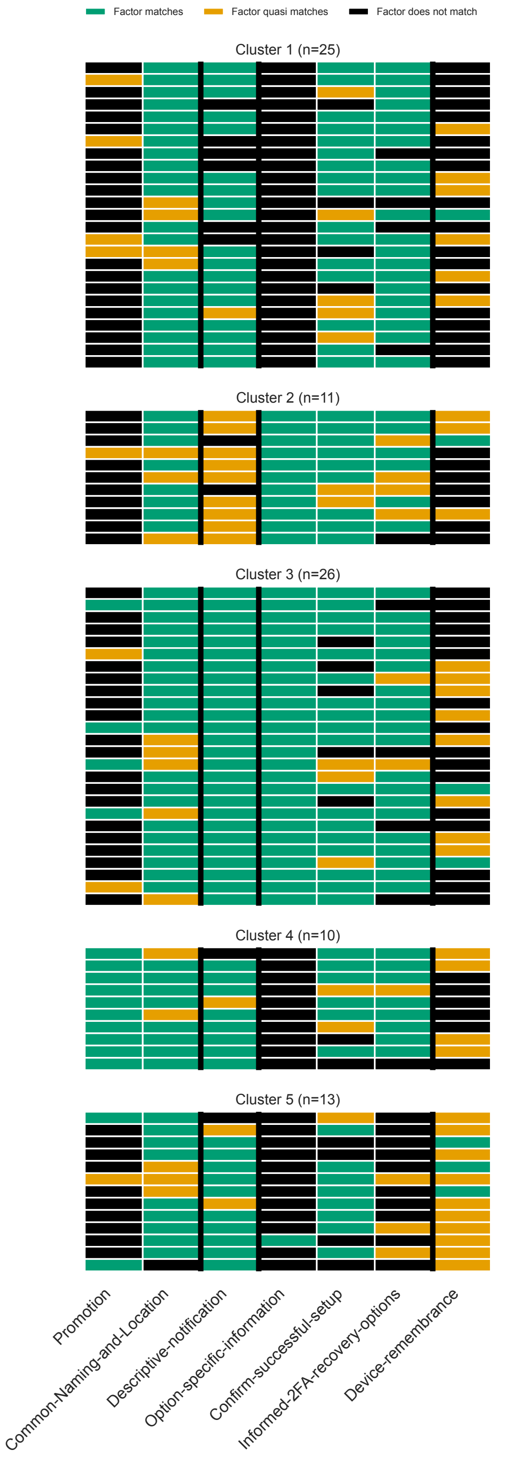


Figure C.14: Clusters based on *Non-conditional-UX* comparison factors.

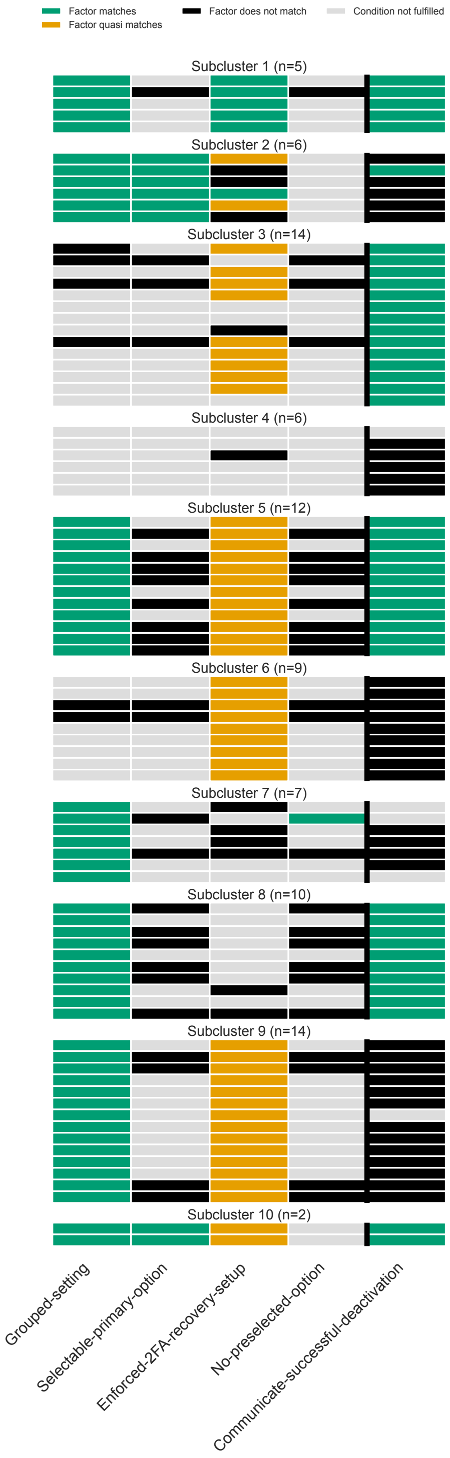


Figure C.15: Clusters based on *Conditional-UX* comparison factors.

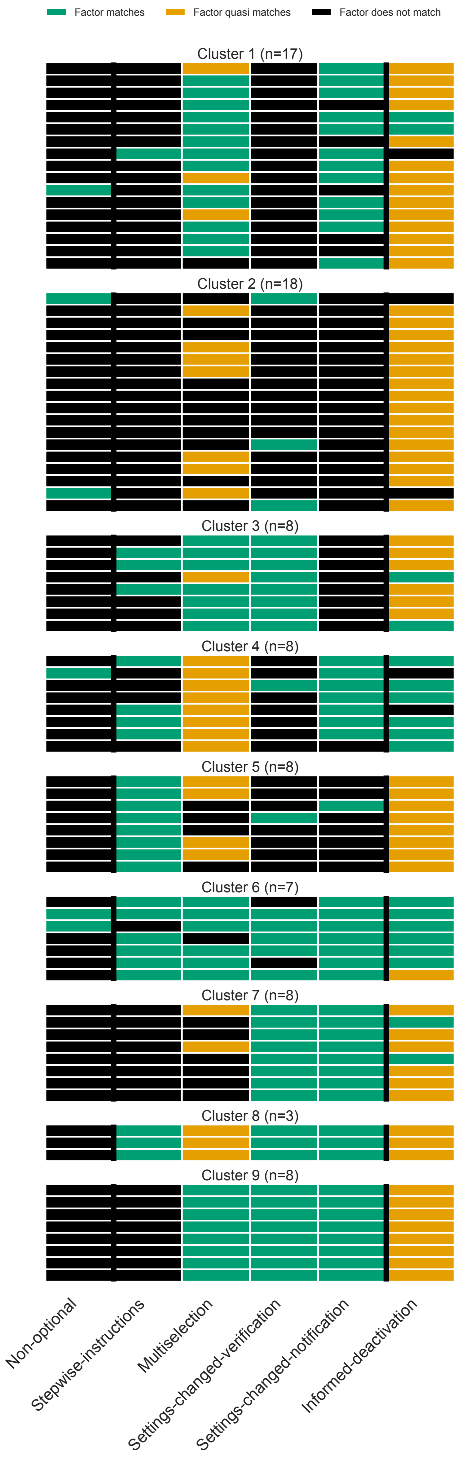


Figure C.16: Clusters based on *Non-conditional-security* comparison factors.

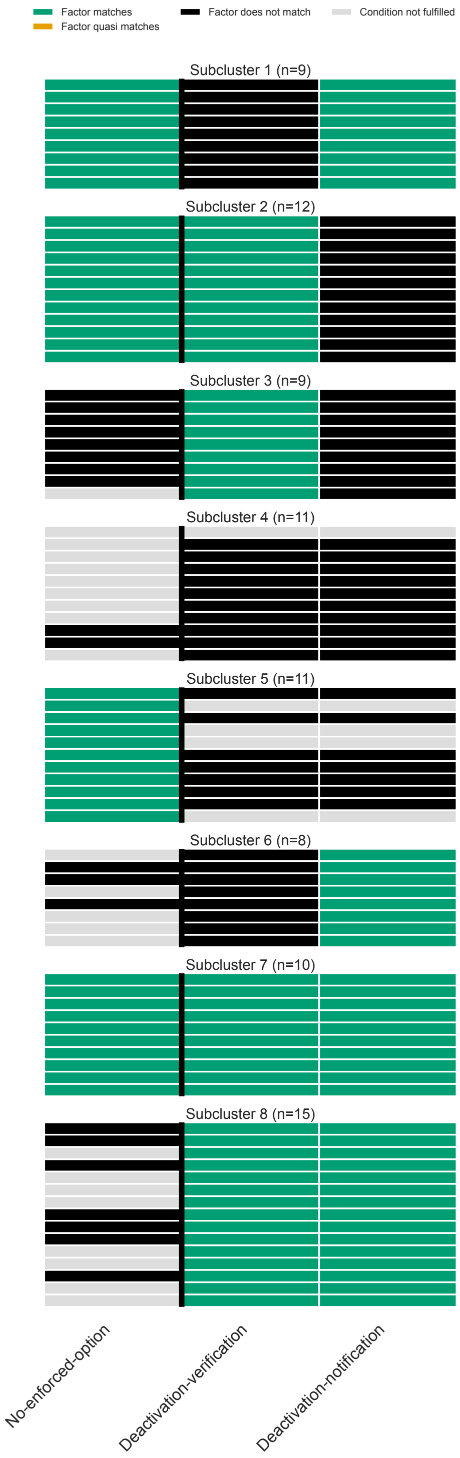


Figure C.17: Clusters based on *Conditional-security* factors.

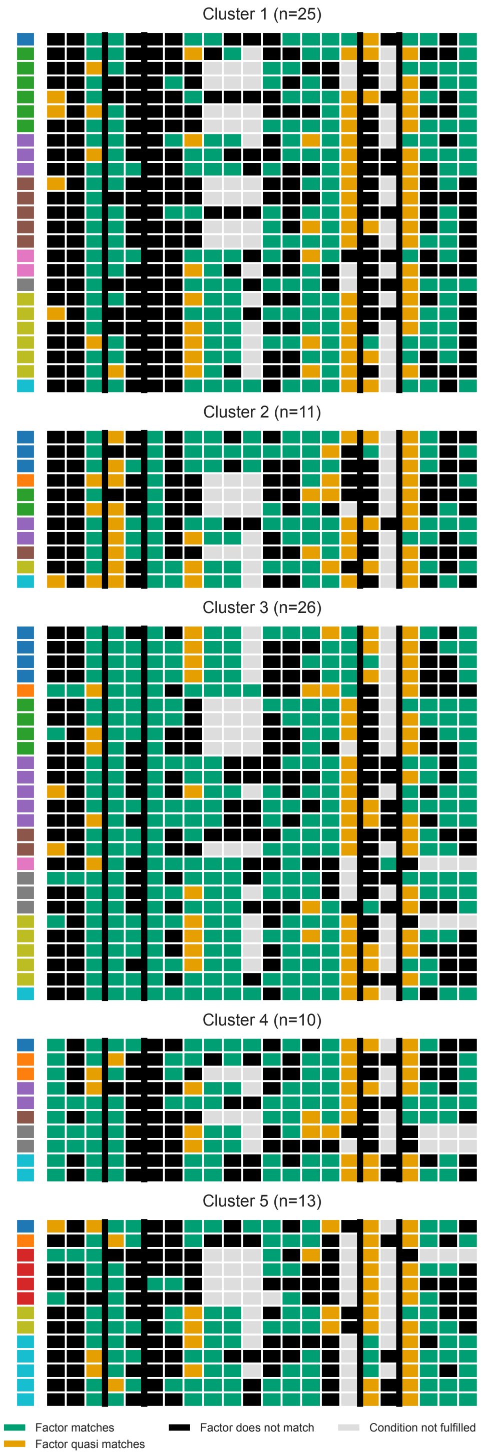


Figure C.18: Clusters based on *all purely UX-related comparison factors*, using non-conditional factors for *inter*-clustering and conditional factors for *intra*-clustering (left column).

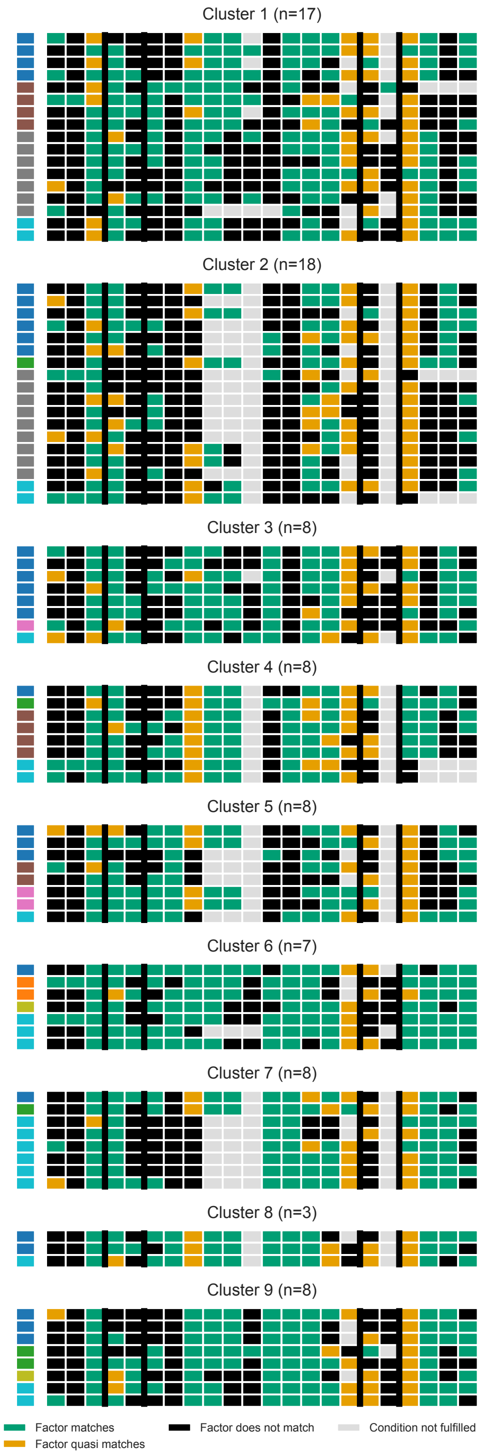


Figure C.19: Clusters based on *all security-related comparison factors*, using non-conditional factors for *inter*-clustering and conditional factors for *intra*-clustering (left column).