

TECHNISCHE
UNIVERSITÄT
DARMSTADT

KNOW YOUR CONTACTS: USABLE AUTHENTICATION FOR SECURE MESSAGING

TOBIAS MÜLLER

Master's Thesis

September 30, 2026

Secure Mobile Networking Lab
Department of Computer Science
Technische Universität Darmstadt

SECURE MOBILE NETWORKING

Tobias Müller: *Know Your Contacts: Usable Authentication for Secure Messaging*, Master's Thesis, Technische Universität Darmstadt, 2026

SEEMOO-MSC-0409

Date of submission: September 30, 2026

Advisor: Prof. Dr.-Ing. Matthias Hollick

Supervisor: Florentin Putz

Secure Mobile Networking Lab
Department of Computer Science
Technische Universität Darmstadt

ABSTRACT

End-to-end encrypted messengers such as Signal commonly rely on servers to distribute users' public keys. A server that also controls message delivery may substitute these keys to perform a machine-in-the-middle attack. Users must therefore authenticate their contacts' public keys through an authentication ceremony to detect such attacks, but prior research identified persistent usability problems in Signal's ceremony. This thesis proposes and evaluates a redesigned ceremony that uses clearer terminology, verification badges, one-way QR code scanning, and an automated Web of Trust approach that propagates verification states. Two pluralistic usability walkthroughs with six participants and three iterations of simplified user testing with 15 participants informed and refined the implementation in the open-source messenger twonly. In the final iteration, all five participants completed the ceremony, and usability issues reported in previous research did not reappear. Its public release enabled the collection of in-app usage statistics from 32 participants, which measured how often contacts were verified and how widely the automation was adopted. A comparative online survey with 23 users of both messengers assessed which elements could be transferred to Signal: 65.2% preferred twonly's verification process, whereas 17.4% preferred Signal's. However, even after the improved ceremony was published, the in-app statistics showed that manual verification barely increased, which suggests that better usability alone does not motivate users to verify their contacts.

ZUSAMMENFASSUNG

Ende-zu-Ende-verschlüsselte Messenger wie Signal verteilen die öffentlichen Schlüssel ihrer Nutzenden über Server. Kontrolliert derselbe Server auch die Nachrichtenzustellung, kann er die Schlüssel austauschen und einen Machine-in-the-Middle-Angriff durchführen. Nutzende müssen die Schlüssel ihrer Kontakte daher über eine Authentifizierungszeremonie authentifizieren, doch frühere Arbeiten zeigten anhaltende Usability-Probleme in Signals Zeremonie. Diese Arbeit entwirft und evaluiert eine überarbeitete Zeremonie mit klarerer Terminologie, Verifizierungsabzeichen, einseitigem Scannen eines QR-Codes und einem automatisierten Web-of-Trust-Ansatz, der Verifizierungszustände weitergibt. Zwei Pluralistic Usability Walkthroughs mit sechs Teilnehmenden und drei Iterationen Simplified User Testing mit 15 Teilnehmenden verfeinerten die Implementierung im quelloffenen Messenger twonly. In der letzten Iteration schlossen alle fünf Teilnehmenden die Zeremonie ab, ohne dass die in früheren Arbeiten berichteten Usability-Probleme erneut auftraten. Die Veröffentlichung ermöglichte In-App-Statistiken von 32 Teilnehmenden, die erfassten, wie häufig Kontakte verifiziert und wie verbreitet die Automatisierung genutzt wurden. Eine Online-Umfrage mit 23 Nutzenden beider Messenger bewertete, welche Elemente auf Signal übertragbar sind: 65,2 % bevorzugten twonlys Verifizierungsprozess, 17,4 % den von Signal. Aber selbst nach der Veröffentlichung der verbesserten Zeremonie stieg die manuelle Verifizierung laut den In-App-Statistiken jedoch kaum an, was darauf hindeutet, dass bessere Usability allein nicht zum Verifizieren motiviert.

Abstract iii
Zusammenfassung iii
List of Figures vi
List of Tables vi
List of Abbreviations vii
1 Introduction 1
2 Background and Related Work 3
2.1 End-to-End Encrypted Messengers . . . 3
2.2 Discount Usability . . . 3
2.3 Usability of State-of-the-Art Authentication Ceremonies . . . 4
3 Research Question 6
4 Methods 7
4.1 Prototyping and Development . . . 7
4.2 Discount Usability Study Design . . . 7
4.3 In-App Survey and Automated Data Collection . . . 9
4.4 Additional User Feedback Collection . . . 9
4.5 Comparative Online Survey . . . 10
4.6 Participants . . . 10
4.7 Ethical Considerations . . . 12
5 Redesign of the Authentication Ceremony 13
5.1 Change in Wording . . . 14
5.2 Motivation Beyond Active MITM Attacks . . . 14
5.3 Fewer and More Intuitive Steps . . . 14
5.4 Web of Trust for Messengers (WoTM) . . . 16
6 Results of the Pluralistic Usability Walkthrough 19
6.1 Iteration 1 - UI Prototype . . . 19
6.2 Iteration 2 - First Implementation . . . 22
6.3 Summary . . . 26
7 Results of the Simplified User Testing 27
7.1 Iteration 3 - First Simplified User Testing . . . 27
7.2 Iteration 4 - Second Simplified User Testing . . . 30
7.3 Iteration 5 - Third Simplified User Testing . . . 32
7.4 Summary . . . 34
8 Results of the Survey and In-App Usage 35
8.1 Survey Demographics and Feature Awareness . . . 35
8.2 Reasons for Verification and Non-Verification . . . 35
8.3 Automated Verification and Transitive Trust . . . 36
8.4 Usability Priorities and Preference . . . 36
8.5 Perceived Security . . . 37
8.6 Terminology . . . 37

8.7 Verification Badges . . . 38
8.8 Summary . . . 38
9 Discussion 39
9.1 Automate Trust Transparently . . . 39
9.2 Reduce the Steps Required for Mutual Verification . . . 39
9.3 Clarify Who Verifies Whom . . . 39
9.4 Address Motivation Beyond Usability . . . 40
9.5 Contextual and Reputation-Based Trust . . . 40
9.6 Provide Instructions at the Point of Action . . . 40
9.7 Provide Sensible Defaults . . . 41
10 Limitations 42
10.1 Implementation in twonly and Transferability to Signal . . . 42
10.2 Security Validation and Trust Assumptions . . . 43
10.3 Exploratory Scope and Sample Size . . . 43
10.4 Participant Bias in In-App and Comparative Online Surveys . . . 43
10.5 Convenience Sampling and ATI Skew . . . 44
10.6 Language and Cultural Bias . . . 44
11 Future Work 45
11.1 Motivation, Incentives, and Gamification . . . 45
11.2 Evaluate Effectiveness Against Phishing . . . 45
11.3 Blue Verification Badge Complexity . . . 45
11.4 Trust Decay and Re-Verification . . . 46
11.5 Integration into Established Messengers . . . 46
12 Conclusion 47
12.1 Availability . . . 48
A Study Policy (Pluralistic Walkthrough) 49
A.1 Guidelines for the whole session . . . 49
A.2 Step-by-step study protocol . . . 49
B Study Policy (Simplified User Testing) 52
B.1 Guidelines for the whole session . . . 52
B.2 Step-by-step study protocol . . . 52
C Tasks (Simplified User Testing) 54
D Task Questionnaire (Simplified User Testing) 56
E Comparative Online Survey 57
E.1 Messenger Usage Comparison . . . 57
E.2 Motivation, Importance, and Preference . . . 57
E.3 twonly-Specific: Verification Badges . . . 58
F Person Questionnaire 60
F.1 Demographic data . . . 60
F.2 Technical knowledge . . . 60
F.3 Technical affinity . . . 61
Bibliography 62

LIST OF FIGURES

Figure 1 Notification flow of a scanned QR code 15
Figure 2 Transitive trust propagation in the WoTM 17
Figure 3 Flow of Key Share Distribution in the WoTM 18
Figure 4 Iteration 1 (UI prototype onboarding) 19
Figure 5 Iteration 1 (UI prototype verification onboarding) 20
Figure 6 Iteration 1 (UI prototype verified and transitive states) 20
Figure 7 Iteration 2 modifications (Onboarding and Mutual Friends Settings) 22
Figure 8 Iteration 2 modifications (In-App Verification) 23
Figure 9 WoTM settings and engagement 24
Figure 10 Verified contacts after the verification badge update 24
Figure 11 Iteration 3 modifications 27
Figure 12 Iteration 4 modifications 30
Figure 13 Iteration 5 modifications 32
Figure 14 Verified contacts by method and week 36

LIST OF TABLES

Table 1 Demographics of all participants 11
Table 2 Iteration 1 demographics 20
Table 3 Iteration 2 demographics 23
Table 4 Demographics and collection modes for additional feedback 25
Table 5 Iteration 3 results 28
Table 6 Iteration 4 results 31
Table 7 Iteration 5 results 33
Table 8 Comparison of verification feature awareness and usage 35
Table 9 Importance of contact verification aspects 37

LIST OF ABBREVIATIONS

ATI	Affinity for Technology Interaction
E2EE	End-To-End-Encryption
GDPR	General Data Protection Regulation
HMAC	Hash-Based Message Authentication Code
MITM	Machine-In-The-Middle
PID	Personally Identifiable Data
PUW	Pluralistic Usability Walkthrough
SUT	Simplified User Testing
WoTM	Web of Trust for Messengers

1 INTRODUCTION

End-to-end encrypted messengers such as WhatsApp and Signal allow billions of users to communicate securely over the internet [1]. They are also increasingly used by high-risk individuals such as journalists [2] and politicians [3] who exchange sensitive information [4]. With the growing adoption of secure messaging, however, governments and malicious actors have an increasing interest in compromising these systems [5, 6].

There are several ways in which a normally secure communication channel can be compromised. Users themselves can make mistakes, as demonstrated by the "SignalGate" incident involving the United States government [4]. Attackers also target journalists and politicians through phishing attacks [2]. In 2026, for example, "presumably Russian state actors" compromised the Signal account of the President of the German Bundestag by impersonating Signal support and tricking the victim into granting device access [7].

Beyond attacks on individual users, more severe attacks can compromise entire messaging systems. The Dutch police, for example, decrypted over 258,000 messages from the end-to-end encrypted messenger IronChat [8], reportedly through an active Machine-In-The-Middle (MITM) attack that replaced encryption keys. Users did not detect the compromise because "warning messages that notified users when their contacts' encryption keys had changed were easy to overlook" [9].

Such MITM attacks are also feasible in both Signal and WhatsApp, since the server controls both public key distribution and message delivery. This control extends beyond the initial key exchange: users of both messengers can reset their keys at any time via phone number verification. The server then distributes the new key to all contacts. In Signal, for contacts whose safety numbers have not been marked as verified, such replacements may be accepted without explicit approval, while a safety-number change notification is displayed. If a safety number has been marked as verified, any change requires manual approval before a new message can be sent [10]. Key-change warnings can be overlooked, as demonstrated by the IronChat incident [9]. A compromised server may therefore substitute public keys, leaving connections that have not been manually verified vulnerable when users overlook the corresponding warnings.

While Key Transparency [11, 12] addresses the risk of active MITM attacks by enabling the verification of public key distribution, it "does not verify real-world identity and does not prevent impersonation or account compromise" [12]. Moreover, Key Transparency is unlikely to improve trust or perceived security and can even make users feel less secure [13].

To defend against these attacks, users can manually perform the authentication ceremony and use a side channel to verify that they possess their contact's correct public key. This verification provides a cryptographic guarantee and helps users establish trust in their contact's real-world identity because scanning QR codes in person physically links the cryptographic key to the device holder. This ceremony could therefore also help prevent phishing and accidental contact confusion by indicating when an identity has not yet been verified. However, authentication ceremonies are rarely performed in practice because current implementations suffer from poor usability [14, 15, 16, 17].

This thesis implements and evaluates a redesigned authentication ceremony based on previous design recommendations, including clearer terminology, verification badges, one-way QR code scanning, and an automated Web of Trust approach named Web of Trust for Messengers (WoTM). Its contribution is evidence from implementing and deploying these ideas in twonly: how transitive verification contributed to verified-contact coverage, how familiar terminology affected users' understanding, and whether improved usability was accompanied by more manual verification. Iterative studies followed Nielsen's qualitative discount usability method [18]. In-app usage statistics and a comparative online survey complemented the lab findings with observations of everyday use and users' preferences.

The remainder of this thesis is structured as follows. Section 2 provides the background and related work. Section 3 formulates the research question, and Section 4 details the methods. Section 5 presents the redesigned ceremony. Section 6 reports the Pluralistic Usability Walkthrough (PUW) findings that informed the implementation, and Section 7 reports its iterative refinement through Simplified User Testing (SUT). Section 8 presents the comparative survey and in-app usage results. Section 9 examines the contributions to prior work before summarizing confirmed design recommendations. Section 10 and Section 11 address limitations and further research. Finally, Section 12 summarizes the findings and answers the research question.

2 BACKGROUND AND RELATED WORK

This chapter introduces the fundamentals of End-To-End-Encryption (E2EE) messengers and the theoretical background of discount usability before reviewing related work on the usability of state-of-the-art authentication ceremonies.

2.1 END-TO-END ENCRYPTED MESSENGERS

Modern instant messaging systems implement E2EE to ensure the confidentiality and integrity of exchanged data. These systems rely on asymmetric cryptography, in which each participant generates a key pair consisting of a private and a public key. Typically, the service provider distributes the public keys through central directory servers.

2.1.1 *Key Establishment and Authentication*

To establish a secure channel, communication partners typically employ the Diffie-Hellman key exchange protocol. Each participant derives a shared secret from their private key and the other participant's public key. This secret enables symmetric encryption of the message payload.

However, this distribution model leaves a security gap when users do not verify the authenticity of public keys. Unless users manually compare safety numbers or fingerprints, the system cannot guarantee that the received public key belongs to the intended recipient rather than to an adversary conducting an active MITM attack [19].

2.1.2 *Phishing Attacks and Contact Confusion*

Phishing attacks typically rely on social engineering rather than protocol vulnerabilities [20]. Attackers may impersonate service personnel to obtain SMS verification codes or registration PINs, allowing them to take over an account without compromising the encryption itself.

Contact confusion can expose communications when users accidentally select an unintended recipient [4]. In this case, encryption still protects the communication channel, but the messages reach someone the sender did not intend to include.

2.2 DISCOUNT USABILITY

Discount usability is a qualitative research approach that focuses on testing rapid prototypes with users [18]. It can yield better results than large-scale usability studies because it emphasizes early and frequent usability input through an iterative process [21]. This thesis used the following two discount usability methods.

2.2.1 *Pluralistic Usability Walkthrough*

Prior to software development or prototype creation, the PUW method [22] can be utilized. The method was originally proposed by Bias [23] and published in Nielsen's book "Usability Inspection Methods" [24]. The goal of the PUW method is to involve representative users who discuss their perspectives on usability during sessions with other users and subsequently with the developer. These sessions often uncover misleading terminology or reveal instances where generally accepted terms for specific concepts have been omitted. During the study, each screen of the new design is subjected to a structured discussion regarding its usability. This process enables immediate redesigns during the session, allowing representative users to provide direct feedback on the proposed changes. The results of these tests are then used to improve the designs before creating a prototype. The PUW method is primarily limited to evaluating terminology and single-screen designs.

2.2.2 *Simplified User Testing*

SUT emphasizes qualitative feedback over quantitative metrics, utilizing the think-aloud method. In this iterative approach, participants independently perform specific tasks. Conducting multiple iterations is recommended, as each cycle typically yields a median usability improvement of 38% [25]. Redesigns addressing specific issues may temporarily decrease other usability metrics, making iterative cycles essential. A sample size of five participants per iteration is recommended [25]; testing additional users yields diminishing returns. During the sessions, participants think aloud to provide direct insights into cognitive obstacles and usability issues as they arise.

2.3 USABILITY OF STATE-OF-THE-ART AUTHENTICATION CEREMONIES

In 2017, Vaziripour et al. [14] evaluated the pairwise authentication ceremonies of Viber, WhatsApp, and Signal by asking participants to verify one another with and without prior knowledge. They found significant usability issues but did not modify the applications themselves. In a follow-up study [26], they proposed UI changes to help users find and complete the ceremony in Signal, showing that discoverability is a key barrier. Despite these changes, fundamental usability problems remained and most users still struggled to complete the ceremony successfully.

Fassl [15] investigated user-centered design for authentication ceremonies in his 2018 master's thesis, using participatory design workshops to create prototypes based on user input. In 2021, Fassl et al. [27] continued this work with collaborative workshops, expert evaluations, and storyboard prototyping. However, their designs were never tested in a real implementation, only through static storyboards. Fassl and Krombholz [17] also conducted an autoethnographic study in 2023, where the first author spent five months trying to authenticate all their contacts. This revealed that even security researchers struggle with authentication ceremonies due to cognitive load, forgetfulness, and the

social awkwardness of asking others to verify. While they identified these issues, they did not test an improved ceremony to address them.

In 2024, Kebbedies [16] confirmed the usability issues discovered by Vaziripour et al. and tried to improve Signal's authentication ceremony through iterative user studies. The success rate of verifications did improve, but several recurring issues could not be resolved, such as users expecting security options in the general app settings, wanting the other user to be notified about a verification request, and being confused by wordings like "safety number" or "coordinator."

Also in 2024, Häring et al. [28] proposed a social authentication protocol as a Signal extension that uses external providers like Facebook for verification. This introduces its own usability problems, as users need to publish their keys on these external platforms.

3

RESEARCH QUESTION

While previous research has identified many usability issues in authentication ceremonies, none of the proposed solutions have fully resolved them. Kebbedies [16] was able to improve the success rate through iterative user studies, but problems like unintuitive wording and poor discoverability of security settings kept coming back across all iterations. Fassl [15, 27] proposed several design recommendations, including approaches like trusted friends, but these were never implemented in an actual messenger where real users could try them out. Whether these ideas would work in practice and how users would perceive them remains an open question.

This thesis therefore addresses the following research question:

RQ: *"How could Signal's authentication ceremony be improved to be more user-friendly?"*

4 METHODS

This chapter details the prototype development, the discount usability study design, the in-app survey, and the comparative online survey. Finally, it summarizes participant demographics and outlines ethical considerations.

4.1 PROTOTYPING AND DEVELOPMENT

The first PUW used the interactive UI prototype shown in Figure 4, created with the open-source platform Penpot [29]. The prototype also included scenario descriptions to guide context transitions, such as moving from onboarding to the friend suggestions view.

The proposed changes in Section 5, especially the automated WoTM, require fundamental modifications to the messenger's architecture, such as contact sharing. Integrating these changes into the original Signal application would have been significantly more complex and would have prevented a public release necessary to collect real-world usage data. The proposed modifications were therefore implemented in the open-source messenger twonly [30], which is maintained by the author of this thesis. Both twonly and Signal use libsignal [31], an implementation of the Signal Protocol. They also rely on centralized, server-controlled public key distribution and message delivery. These shared protocol and architectural properties impose the same requirements for preventing active MITM attacks on both authentication ceremonies and therefore enable a meaningful comparison between their implementations.

4.2 DISCOUNT USABILITY STUDY DESIGN

To iteratively improve and evaluate the authentication ceremonies proposed in Section 5, discount usability [18] was employed (see Section 2.2), beginning with an initial PUW and continuing with iterative SUT sessions. Both study types shared the same overall structure, differing primarily in the study tasks and the moderator's interaction with participants. The corresponding study policies are provided in the appendix under Section A and Section B. The setup of the SUT closely followed the design from Kebbedies [16], and the sessions were conducted in quiet, office-like environments. Before starting, participants were asked if any room elements were distracting them and if they consented to audio recording for transcription.

For both study types, the moderator followed the step-by-step protocols in Section A.2 and Section B.2, which covered obtaining General Data Protection Regulation (GDPR) consent, delivering introductory remarks, and guiding interactions with participants.

4.2.1 *Pluralistic Usability Walkthrough*

The PUW consisted of a group session with three representative users and the moderator. A sample size of three was selected as a compromise between resource constraints and

evaluation depth, which is sufficient to uncover approximately 60% to 65% of usability issues [32]. The original PUW method requires different individuals to serve as researcher and developer, but resource constraints required the moderator to assume both roles. Participants recorded their initial responses silently before each group discussion to minimize observer bias.

The first PUW iteration utilized the UI prototype described in Section 4.1. The second iteration used screenshots of the actual implementation embedded in a presentation along with the tasks and scenarios.

4.2.2 *Simplified User Testing*

A sample size of five participants per iteration was targeted, following Nielsen's recommendation [25]. Each SUT session involved a single participant and the moderator. Participants used the think-aloud method to explain their cognitive obstacles.

In the SUT, participants performed two or three tasks (detailed in Section C) that evaluated how users perceived WoTM and whether they could complete the authentication ceremonies.

Sessions were terminated if a participant stayed on an incorrect path for over ten minutes. The participant received hints if a task failed. Afterward, participants answered open-ended questions about the functionality, completed the questionnaire in Section F to record demographics and control variables, received cash compensation, and signed a receipt.

4.2.3 *Quantitative Data*

The post-task questionnaire in Section D contained six statements on a 5-point Likert scale to provide quantitative data on perceived usability. This matched the design by Kebbedies [16] and incorporated metrics from Vaziripour et al. [14, 26]. Participants also rated how likely they were to recommend the application on a scale from 0 to 10. Due to the small sample size, the questionnaire aimed to identify qualitative tendencies rather than statistically significant results. By asking participants to explain their choices, the questionnaire helped gather qualitative feedback.

Following Kebbedies [16], the study tracked completion time and task success (pass). Completion time was measured from the start instruction ('you may now start') until the participant declared the task solved. A pass was achieved when the user interface indicated that the authentication ceremony had been completed for both users.

The questionnaire responses, performance metrics, demographics, and control variables were processed using a Python script. To compare the quantitative results of each iteration, a transformation proposed by Nielsen [25] was applied to the Likert-scale answers. This transformation yields comparable relative usability scores: it maps the answers to a range of $[-1, 1]$, applies the arcsine function to weight extremes, and uses the exponential function e^x to stretch the scale, ensuring consistent percentage improvements. For each subsequent SUT iteration, percentage improvements were calculated relative to the previous iteration.

4.2.4 *Qualitative Data*

Transcripts from the think-aloud sessions, walkthrough discussions, and interviews were analyzed. The analysis focused on participants' direct feedback and suggestions for improvement. For unsuccessful tasks, participants' actions and explanations were examined to identify the reasons for failure. Depending on their actions or if a task failed, participants were asked additional questions to understand the reasoning behind their decisions. These insights helped explain the quantitative results and guided modifications in subsequent iterations.

4.2.5 *Control Variables*

Control variables were included to account for prior technical experience. Self-reported metrics assessed familiarity with everyday technologies and awareness of IT security. Technology affinity was measured using the short version of the Affinity for Technology Interaction (ATI) scale developed by Wessel et al. [33].

4.3 IN-APP SURVEY AND AUTOMATED DATA COLLECTION

To supplement the results of the SUT and PUW with observational data on the adoption and use of WoTM, an in-app survey was conducted at the start of the study. The survey requested permission to collect usage statistics in the background to determine whether users accepted and enabled the WoTM approach. These data also showed whether and how users verified their contacts and how the overall number of verifications changed after the improved authentication ceremony was released.

The application collected and uploaded usage statistics only when a participant opened it. A participant was therefore included in a weekly mean only if a record was available from that calendar week. If more than one record was available, only the most recent was used. Values from previous weeks were not carried forward, because doing so would treat an earlier state as a current observation and obscure changes in the weekly sample. For each participant, the share of contacts verified through each method was calculated. These individual shares were then averaged to obtain the weekly values. As a result, each participant contributed equally to the weekly mean, regardless of how many contacts they had. The participant count shown in Figure 14 can therefore fall when previously observed participants do not open the application during a given week.

4.4 ADDITIONAL USER FEEDBACK COLLECTION

After changes from an iteration were published, active users were invited to provide feedback by email or through the contact form in the application. Other users were approached in person. No fixed questionnaire or interview guide was used. Users could describe their impressions and suggestions in their own words and voluntarily provide information about their age and self-assessed security knowledge. This resulted in an unstructured collection of feedback from eight users. The comments were reviewed to

identify usability problems and suggestions for subsequent iterations. This feedback was accounted for separately from the three structured studies and was not added to the thesis-wide participant total.

4.5 COMPARATIVE ONLINE SURVEY

After the version resulting from the final SUT had been available in twonly for several weeks, the comparative online survey was conducted anonymously among users who had actively used both twonly and Signal for more than one month. A self-administered format was chosen to minimize social desirability bias [34] and eliminate the observer effect [35]. The survey addressed RQ by comparing the final design with Signal's implementation to determine which of the proposed changes could be transferred to Signal. Furthermore, the survey assessed users' satisfaction with the improved authentication ceremony, whether it changed their usage behavior, and whether its increased prominence encouraged more active engagement.

4.6 PARTICIPANTS

Between January and August 2026, 32 participants joined the in-app study by completing the survey, which enabled repeated collection of their verification usage when they opened the application. The survey was advertised with a public Mastodon post to encourage participation [36]. As Mastodon tends to attract a more technically oriented user base, around 62% of participants reported advanced computer knowledge, with 38% reporting advanced IT-security knowledge.

Between April 2026 and July 2026, 23 participants were recruited for the usability studies. Two participants completed pilot studies (one for each study type) and were excluded from the final data, leaving a sample of 21: six for the PUW and 15 for the individual SUT sessions. Recruitment occurred primarily through acquaintances and in-person at the university. To achieve a balanced sample, participants were pre-screened for security knowledge, ensuring each iteration included both non-technical and technical users. Participants received €15 for the PUW sessions and €10 for the SUT sessions.

In August 2026, 23 participants completed the comparative online survey, which was again advertised publicly on Mastodon [37]. Similar to the in-app survey, the majority of respondents reported advanced computer knowledge (70%) and advanced IT-security knowledge (65%).

The studies comprised 76 participations. The number of distinct individuals is unknown because both the in-app statistics and the comparative online survey were anonymous, preventing reliable identification of overlap between studies. Table 1 summarizes the demographics for each study.

Table 1: This table summarizes participant demographics for the PUW and SUT, the in-app survey, and the comparative online survey. Counts and percentages are reported separately for each study.

	Variable	**PUW & SUT (n=21)**	**In-App (n=32)**	**Online (n=23)**
Age	18-22	1 (5%)	6 (19%)	2 (9%)
	23-27	13 (62%)	15 (47%)	11 (48%)
	28-32	3 (14%)	4 (12%)	3 (13%)
	33-37	1 (5%)	1 (3%)	3 (13%)
	38-42	0 (0%)	1 (3%)	2 (9%)
	43-47	1 (5%)	0 (0%)	0 (0%)
	48-52	0 (0%)	2 (6%)	0 (0%)
	53-57	2 (10%)	1 (3%)	1 (4%)
	58-62	0 (0%)	1 (3%)	1 (4%)
Smartphone exp.	>2 years	21 (100%)	30 (94%)	23 (100%)
Computer knowledge	Beginner	2 (10%)	2 (6%)	3 (13%)
	Intermediate	10 (48%)	8 (25%)	4 (17%)
	Advanced	9 (43%)	20 (62%)	16 (70%)
IT-security knowledge	Beginner	6 (29%)	9 (28%)	6 (26%)
	Intermediate	9 (43%)	9 (28%)	2 (9%)
	Advanced	6 (29%)	12 (38%)	15 (65%)
Educational degree	Still in school	0 (0%)	2 (6%)	0 (0%)
	Secondary school	0 (0%)	3 (9%)	0 (0%)
	Intermediate secondary	1 (5%)	0 (0%)	1 (4%)
	University entrance qualification	20 (95%)	26 (81%)	22 (96%)
Vocational degree	None	3 (14%)	5 (16%)	3 (13%)
	Apprenticeship	3 (14%)	0 (0%)	2 (9%)
	Technical school	0 (0%)	1 (3%)	0 (0%)
	Bachelor	8 (38%)	9 (28%)	8 (35%)
	Master	4 (19%)	7 (22%)	7 (30%)
	Diplom	2 (10%)	3 (9%)	3 (13%)

	Variable	PUW & SUT (n=21)	In-App (n=32)	Online (n=23)
	Staatsexamen	1 (5%)	0 (0%)	0 (0%)
Student	Yes	12 (57%)	14 (44%)	10 (43%)
	No	9 (43%)	16 (50%)	13 (57%)
ATI	Median	3.33	-	4.56
	IQR	2.4	-	2.78

4.7 ETHICAL CONSIDERATIONS

The study was approved by the Institutional Review Board of the Technical University of Darmstadt. All Personally Identifiable Data (PID) were processed and stored in strict compliance with the GDPR. Individuals signed consent forms explaining the data handling procedures before participating. Participants could opt out of providing demographic details or being audio and screen recorded.

Each individual received a unique five-character alphanumeric ID to protect their privacy. This identifier labeled all their materials, including consent forms, questionnaires, receipts, recordings, and transcripts. All PIDs were stored locally on a laptop using full-disk encryption and backed up on an encrypted external drive.

Audio recordings were transcribed locally using the open-source tool noScribe [38]. Transcripts were then manually corrected and pseudonymized, redacting all PIDs. Upon completing a transcript, the corresponding audio recording was deleted. Physical documents were destroyed once the data were digitized and pseudonymized.

5 REDESIGN OF THE AUTHENTICATION CEREMONY

The authentication ceremony currently implemented in Signal has been shown to suffer from poor usability [14, 16, 26]. To improve usability, prior research primarily focused on increasing the visibility of the unverified state, for example by displaying a red action banner to signal that the current chat was insecure. However, the core concept remained unchanged: verification still occurred on a per-chat or per-group basis, either by comparing safety numbers or through alternative methods such as PairSonic [39]. While this visual highlighting increased the rate of successful verifications during lab studies, the approach still exhibited recurring usability problems [16]:

- Complex or unfamiliar wording, such as “safety number” or “coordinator”
- Security options being expected within the general app settings
- Users wanting the other party to be informed about an initiated verification
- Unclear instructions on how to scan a QR code
- Verification taking longer than expected

To solve these issues, this thesis designs and implements a new way to perform the authentication ceremony based on design recommendations from prior research:

- Transparent about the current state, but not verbose [16]
- Abstract away from cryptographic concepts and unfamiliar wording [16]
- Reduce the steps users need to perform to authenticate each other, making the process as intuitive as possible [16]
- Use recognizable trust indicators, drawing on the color-coded contact states proposed by Fassl et al. [27]
- Automate the process of establishing trust whenever possible [15]
- To motivate users, they need to have some (not necessarily correct) understanding of its purpose and its effects [15]

Our proposed solution introduces five fundamental changes to the current implementation: a change in wording, additional motivation, a reduction in steps, the notification of the scanned contact, and the use of an automated Web of Trust approach.

5.1 CHANGE IN WORDING

Kebbedies' user studies identified cryptographic terminology as a recurring barrier [16]. The redesigned interface therefore uses the term "verification badge" instead of "safety number". Its color-coded states build on the contact trust indicators proposed by Fassl et al. [27]. Familiarity with badges on social media may help users recognize an identity check, but here users verify their contacts themselves instead of relying on the platform. The badge also makes the verification state visible. Whether it helps prevent phishing attacks, such as those reported for Signal [40], requires separate evaluation.

5.2 MOTIVATION BEYOND ACTIVE MITM ATTACKS

Most users neither understand the danger of an active MITM attack nor consider it relevant to themselves, as Kebbedies pointed out: "To add to the point about the threat model, many participants said that securing their information from third parties is not a priority" [16]. Therefore, the changed terminology in Section 5.1 could help users more easily understand the purpose of the verification badge. The concept is known from other platforms, and phishing is a well-known attack to which many people are sensitized by their employers or the news. Bypassing the need to explain complex threat models with the risk of phishing attacks could reduce the social awkwardness associated with the authentication ceremony, an issue raised by Fassl [17].

5.3 FEWER AND MORE INTUITIVE STEPS

One of the key takeaways from Kebbedies' work was the need to reduce the number of verification steps and make the process as intuitive as possible. To address these usability issues, we implement the following changes.

5.3.1 *Availability in the Main Settings and Profile View*

The biggest hurdle for currently implemented systems, and the main reason for the recurring expectation that verification should take place in general settings, is the requirement for verification per chat. In this process, "safety numbers" are compared, which consist of a concatenation of both users' public keys. Due to this concatenation, other users cannot use the same QR code. It also requires the user to open the other person's private chat separately. From a cryptographic perspective, however, it is sufficient for Alice to scan Bob's public key and confirm that it matches the identity key used in their session. Successful authenticated decryption of a subsequent message from Bob confirms that the session was established using Alice's corresponding public key. By moving away from comparing the concatenation and instead verifying only the individual public key, the verification process can be radically simplified. In the system settings, a single QR code can be displayed that anyone can scan without having to navigate to a specific conversation.

5.3.2 *Use QR Code Scanning for the First Contact Exchange*

As a result of the simplification described in Section 5.3.1, the initial contact exchange via a QR code can now easily be used to simultaneously verify the users. By exchanging the public key locally through the QR code instead of retrieving it from a server, the application automatically marks new sessions as verified.

5.3.3 *Notification of the Scanned Contact*

Notifying the scanned contact addresses a recurring expectation identified in Kebbedies' studies and cuts the required verification steps, because only one person has to scan the QR code. When a user scans a QR code, an automated notification is transmitted via the established end-to-end encryption. Figure 1 illustrates how the notification enables both users to verify their connection simultaneously without a mutual scan.

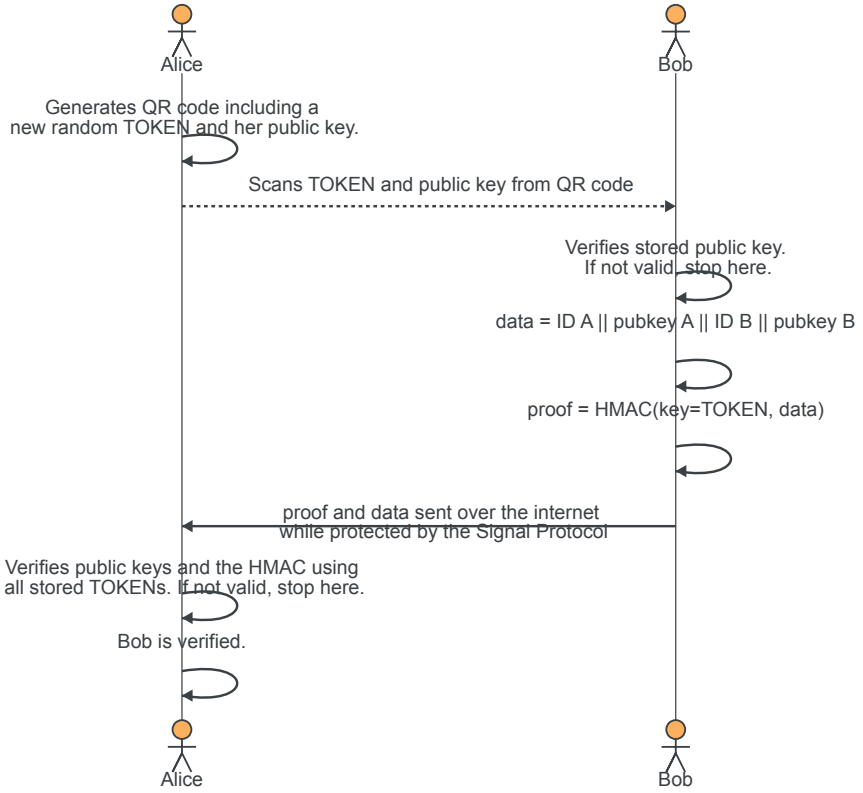


Figure 1: This figure illustrates the one-way notification scan used in the asymmetric authentication ceremony. A secret token embedded in the QR code serves as an HMAC key to authenticate the notification. This ensures the token is never transmitted over the network, preventing active MITM forgery.

To prevent an active MITM attacker from forging this notification, the protocol relies on a secret token, as detailed in Figure 1. The process generates a QR code prefixed with a deep or universal link [41, 42]. Standard scanners or the system camera open the chat

application directly, even if the user scans the QR code with another app. The QR code contains the user's unique ID, public key, signed pre-key, and username so they can create a new session.

The QR code also holds a randomly generated secret token. Because this token is only shared during a physical scan, an active MITM attacker would need physical access to the users or their devices to compromise the ceremony. The token remains valid for one day, allowing verifications to finish in the background if the scan happens offline.

To securely notify the scanned person, the token acts only as an encryption key for an Hash-Based Message Authentication Code (HMAC) calculation over the session's fingerprint. The token never leaves the devices, so a malicious server executing an active MITM attack cannot intercept the token and falsely verify a session. Alice uses her stored tokens to validate the received HMAC, which confirms that Bob has scanned a newly generated token. Alice marks the session as verified after checking that the public keys in the notification match the ones used for decryption.

5.3.4 *Verification through a Shared Profile Link*

Users can also share their own profile link through another channel. The link contains their username and public identity key. When the recipient opens it, twonly compares this key with the key stored for the contact and asks the recipient to confirm that the link came directly from that person before marking the contact as verified. This differs from manual contact sharing in Section 5.4.1: the recipient verifies the person named in a link received directly from that person, whereas manual contact sharing forwards a third person's verified key through an already verified contact.

5.4 WEB OF TRUST FOR MESSENGERS (WOTM)

Fassl [15] described automatic trust through friends as a way to reduce the effort and social discomfort of manual verification. The WoTM implements this concept by propagating verification states through verified contacts, as shown in Figure 2.

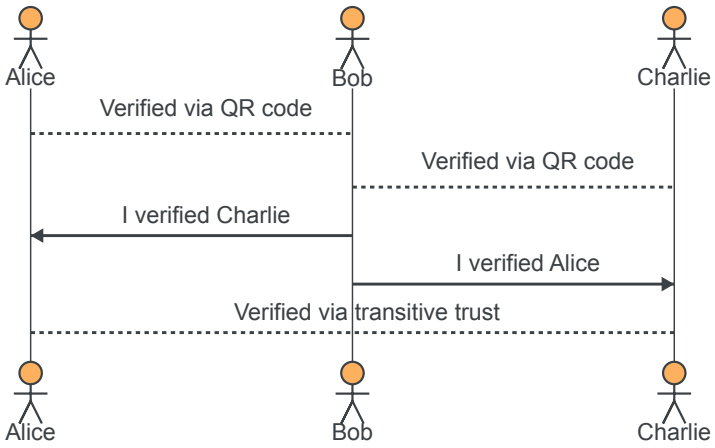


Figure 2: Bob uses his previously verified sessions with Alice and Charlie to securely inform them that he has manually verified their respective public keys. This process allows Alice and Charlie to establish a verified connection through transitive trust without requiring an in-person meeting.

Under this WoTM approach, contacts propagate their verification states to other verified contacts. As shown in Figure 2, this mechanism enables Alice and Charlie to establish a trusted connection without having met in person. Bob could lie about the verification state, so this transitive trust gets a distinct badge color to show it is less confident than a manual QR code scan. The transitive verification state is only passed on for contacts who were verified in person and who meet other requirements, like having received at least four pictures from the user. This requirement obligates a malicious contact to first meet the target in person and then lure them into sending at least four pictures.

To prevent users from revealing their entire social graph, each shared contact entry is encrypted with a key provided by that contact. Consequently, a recipient can only decrypt the contact if they are already connected with that contact and have received their decryption keys. This process is illustrated in Figure 3. Bob and Charlie can decrypt each other in Alice's announcement because they have exchanged their decryption keys. However, they cannot decrypt David because they have not received his keys. Similarly, David is unable to decrypt any of Alice's contacts since he is not directly connected with them.

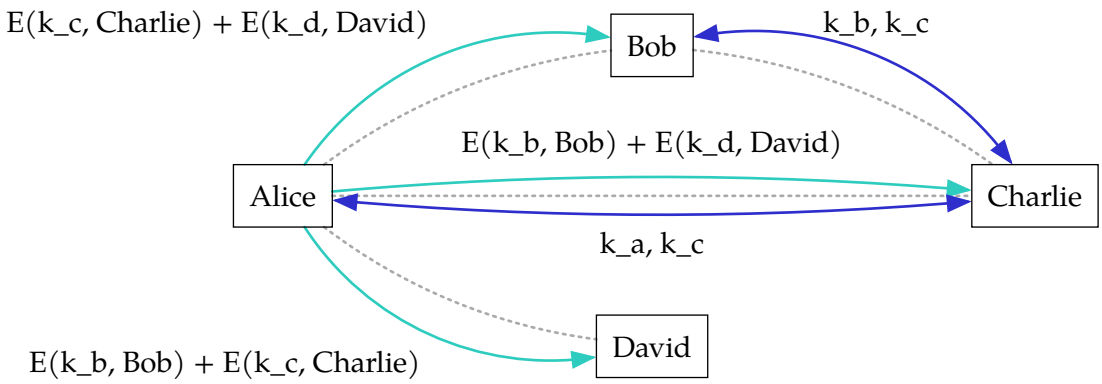


Figure 3: Dotted lines indicate friendships for which WoTM is enabled. Alice shares her individually encrypted contact entries with Bob, Charlie, and David. Since Bob and Charlie are also connected and have already exchanged their decryption keys, they can decrypt each other's details in Alice's announcement. Consequently, each learns that Alice has verified the other and can display this verification state in the UI. Although David also receives Alice's contact entries, he cannot decrypt their details because he has no direct connection with the respective contacts.

5.4.1 *Manual Sharing of Contacts Including Their Verification State*

Beyond the automated WoTM detailed in Section 5.4, the verification badge is used when users manually share contacts. When a user shares a contact with a verified contact, the shared contact's verified public key is included. The recipient's application compares this public key with the version retrieved from the server. The application displays the verification badge for the shared contact if the keys match and the sender is verified. This dual-check prevents a malicious friend or a compromised server from distributing a fraudulent public key.

6 RESULTS OF THE PLURALISTIC USABILITY WALKTHROUGH

This chapter presents the results gathered during the PUWs discussion rounds using the UI prototype and the first implemented version. For each iteration, the chapter describes the modifications made, participant demographics, the quantitative data collected and corresponding improvements, and qualitative data on the usability issues encountered by participants. The results of each iteration are discussed separately, while overarching topics are addressed in Section 9.

6.1 ITERATION 1 - UI PROTOTYPE

The first iteration used an interactive user interface prototype to investigate how users perceived the proposed WoTM implementation and its requirement to share contact lists. The walkthrough also aimed to identify initial usability problems and improve the terminology.

The first UI prototype focused on the onboarding process, which is critical because sharing contact lists for the WoTM approach requires user consent. Figures 4, 5, and 6 show the screens tested in this session and their navigation paths.

During the session, the participants performed several tasks. First, they walked through the onboarding screens in Figure 4 to discuss the configuration options. And then, the participants completed the verification flow in Figure 5 and Figure 6, which involved checking and verifying an unverified contact, and then evaluating a contact verified through the WoTM approach.

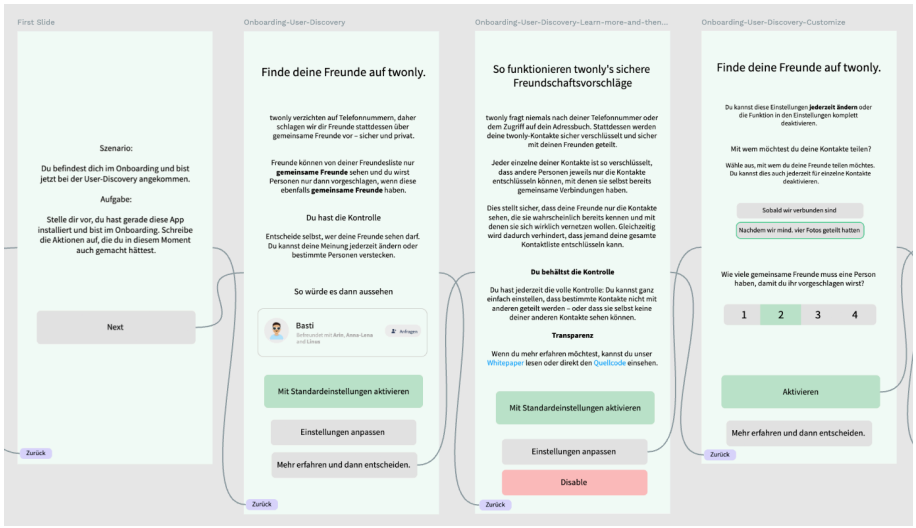


Figure 4: The first iteration presented the onboarding flow and contact-sharing options for the WoTM.

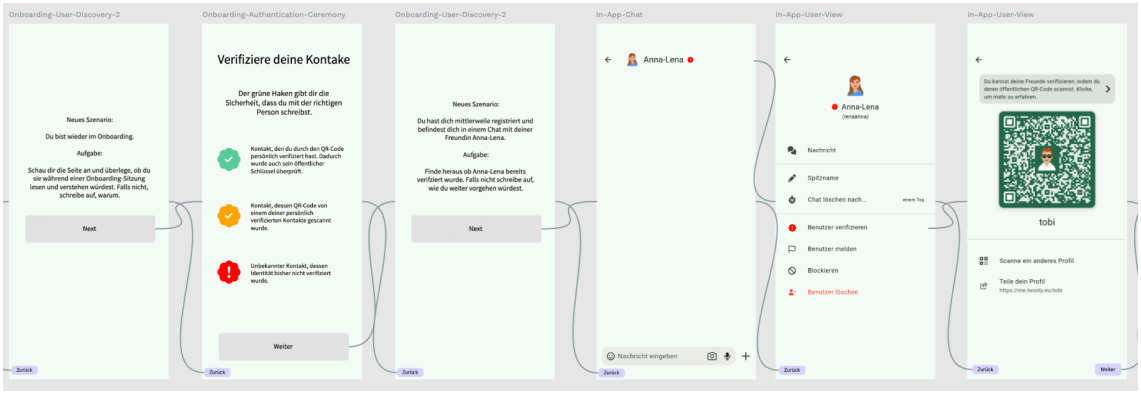


Figure 5: The onboarding screen explaining the verification levels and the interface for an unverified contact.

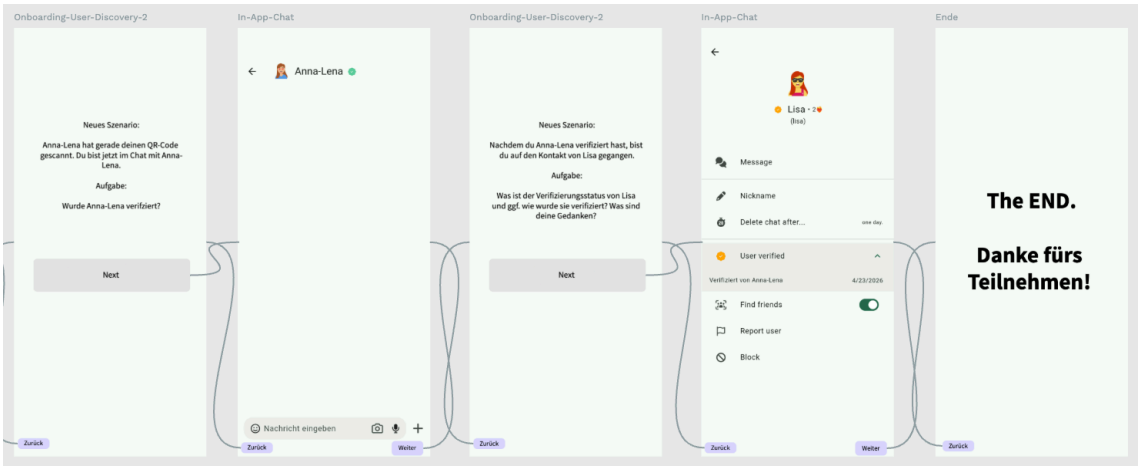


Figure 6: The user interface for a manually verified contact and a verified contact via the WoTM approach.

6.1.1 *Demographics*

The PUW was conducted on 24 April 2026 with the three participants detailed in Table 2, lasting approximately 90 minutes.

Table 2: This table shows a demographic summary of the participants of iteration 1.

ID	Age	Highest Degree	Student	ATI	Security Knowledge
TK4O2	23-27	Apprenticeship	Yes	3.78	Intermediate
A7HK3	23-27	Uni. entrance qualification	Yes	1.22	Intermediate
VKH82	23-27	Apprenticeship	No	1.89	Beginner

6.1.2 *Usability Problems*

The most significant issues identified during the session are grouped into the following categories. Minor observations, such as requests to use bolder headings, were omitted.

1. General Problems
    - **UP-1**: Option for manual approval is requested before a contact is shared with friends (A7HK3)
    - **UP-2**: When clicking on verify user, the explanation of the badge colors should be shown again instead of directly opening the QR code scanner (VKH82)
2. Visual Problems
    - **UP-3**: Important text should be highlighted with the same color as the icons on the verification badge explanation screen (TK4O2)
    - **UP-4**: More subheadings and text highlighting are needed for important parts of the text (TK4O2)
    - **UP-5**: Orange is a misleading color for the verification badge, as it looks like an error even though the contact is verified (VKH82)
3. Text Problems
    - **UP-6**: Onboarding has too much text, requiring more visual examples or animations (A7HK3)
    - **UP-7**: Text is too technical, using complex terms like "whitepaper" or "public key" (VKH82)

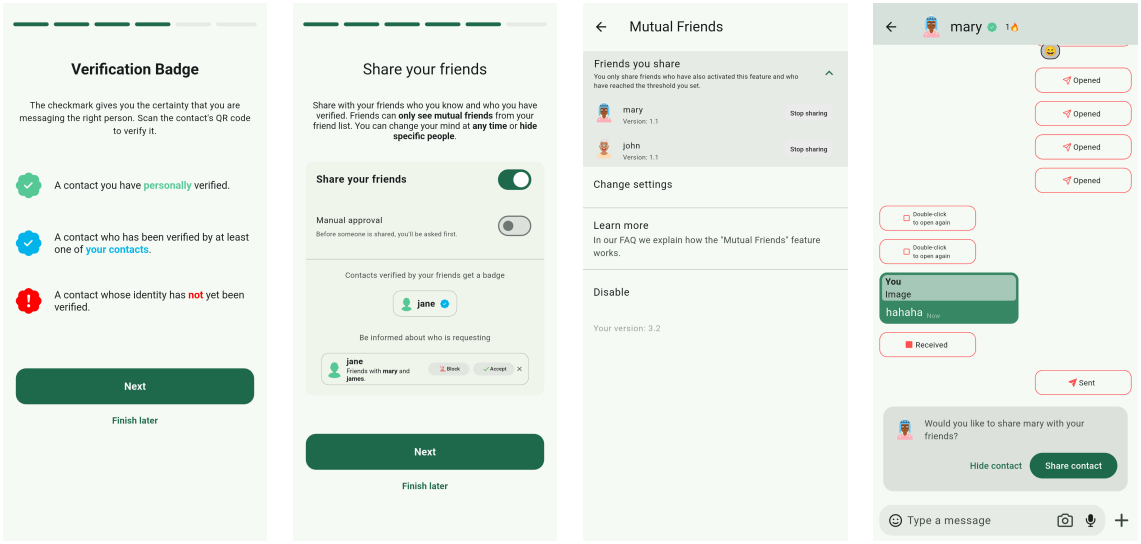
6.1.3 *Discussion*

Despite the usability problems in the interface, all participants supported sharing contacts to propagate transitive trust. Overall, the WoTM approach was considered helpful and could increase trust in the messenger. Below are the key insights from this first iteration.

1. **Visual examples are better than text**: Participants preferred visual examples or animations over long explanations. They also suggested changing the transitive verification badge color from orange to blue to match verified profile indicators on other social media platforms.
2. **WoTM should offer manual approval**: Participants appreciated the ability to hide specific contacts, but emphasized that an option for manual approval, which prompts users before sharing a contact, would be better.

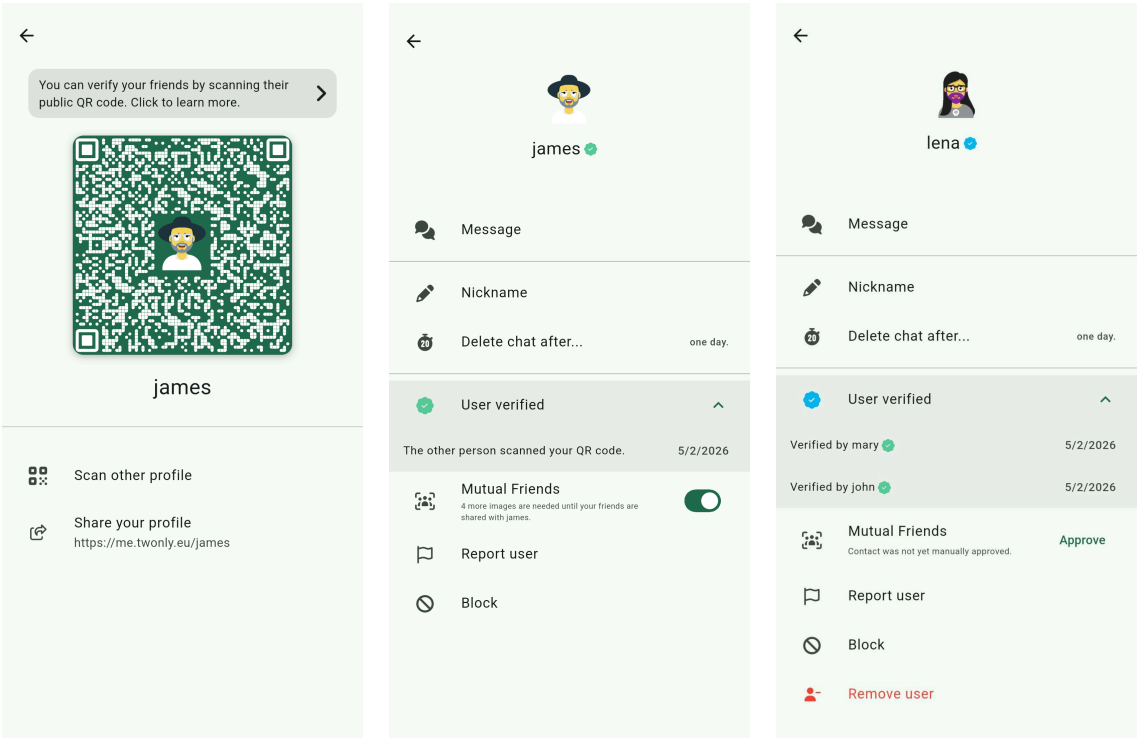
6.2 ITERATION 2 - FIRST IMPLEMENTATION

For the second iteration, the focus shifted from the UI prototype to the first functional implementation with an initial public release [43]. This approach allowed the collection of initial feedback from the participants of the in-app user study. This feedback provided deeper insights into how users interact with the proposed WoTM approach in a real-world scenario rather than a lab setting, where their decisions might differ. For the second iteration of the PUW, three participants took part again, as detailed in Table 3. In contrast to the first iteration, screenshots embedded in a presentation format were utilized. We made this change because participants in the first iteration often advanced to the next screen despite instructions not to, and frequently had to navigate back when they forgot the task or scenario. The presentation format gave us better control over the displayed screen and placed task instructions next to the screenshots, simplifying the process. In total, 13 screenshots were shown to the participants, with the key modifications illustrated in Figure 7 and Figure 8.



(a) (b) (c) (d)

Figure 7: This figure shows the onboarding flow and mutual friends settings. Screen (a) explains the verification badge, and screen (b) presents the onboarding screen to share friends. Screen (c) displays the mutual friends settings page, while screen (d) shows the prompt for manual approval of shared contacts.



(e) (f) (g)

Figure 8: This figure shows the updated in-app verification screens. Screen (e) displays the verification QR codes. Screen (f) shows the interface after a contact has scanned the user's QR code. Screen (g) demonstrates the new blue verification badge.

The second iteration aimed to address all usability issues identified during the first PUW. While most adjustments were minor, we introduced two major modifications:

First, we added clear configuration settings for the WoTM approach to the onboarding flow. Specifically, we presented the option to "share your friends" (b). Along with fixing usability issues UP-4, UP-6, and UP-7, this change led to the new onboarding screen (b). Unlike the first iteration, participants had much less text to read and saw concrete examples that directly illustrated the benefits of enabling the WoTM features.

The second major modification introduced an option for manual approval, as suggested in UP-1. This option could be enabled during the onboarding step on screen (b); if activated, the interface prompted the user for confirmation before sharing a contact (d).

6.2.1 *Demographics*

Table 3: This table shows a demographic summary of the participants of iteration 2.

ID	Age	Highest Degree	Student	ATI	Security Knowledge
K32BL	23-27	Bachelor	Yes	4.33	Advanced
LKM36	23-27	Bachelor	Yes	4.78	Advanced
S5ABC	23-27	Bachelor	No	2.56	Intermediate

The second PUW was conducted on 4 May 2026 and included the three participants detailed in Table 3, with each session lasting approximately 60 minutes.

6.2.2 *In-App Statistics and Additional Feedback*

In addition to the PUW, the new version was released as an update on 1 May 2026 and was subsequently installed by users over the following weeks, depending on their device settings. At the time of the release, 19 participants in the in-app user study were active, which enabled the collection of the in-app statistics shown in Figure 9 and Figure 10.

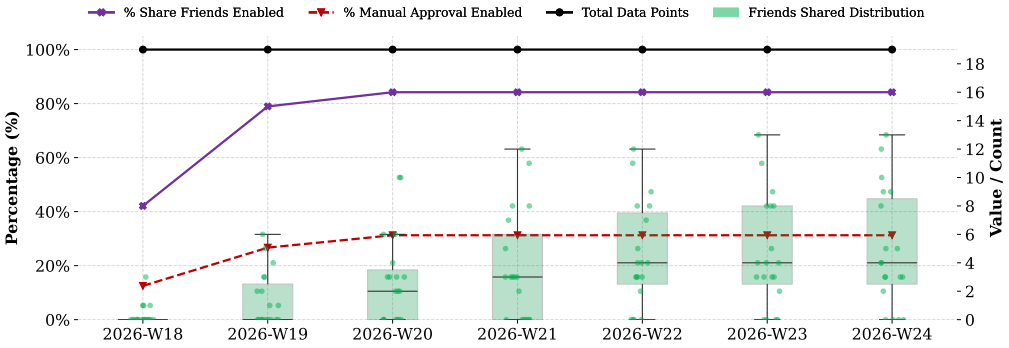


Figure 9: This figure shows the users' chosen WoTM settings and engagement metrics of the participants of the in-app user study. Around 80% had already enabled the feature, with approximately a third having enabled manual approval. Additionally, the boxplot illustrates the distribution of friends shared by each participant.

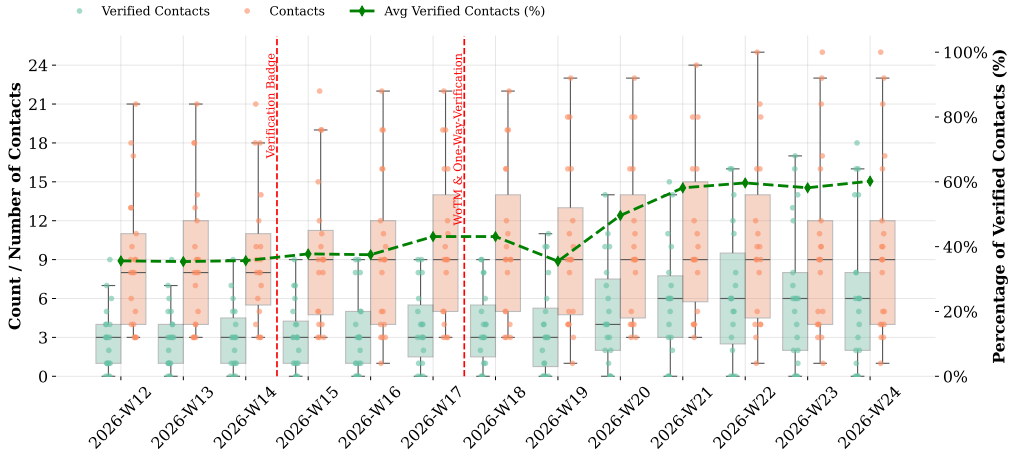


Figure 10: This figure illustrates the evolution of contact verification among participants. Following the introduction of the red verification badge, the median number of verified contacts remained stable and the manual verification rate remained low. The subsequent automated trusted friend update coincided with increases in both the average and median values, while the manual verification remained limited.

The demographics and collection modes of the eight users who provided additional feedback are summarized in Table 4.

Table 4: This table presents the demographics and collection modes of the users who provided additional feedback.

ID	Age	Highest Degree	Student	Security Knowledge	Source
USER_1	18-23	Middle School	No	Beginner	In person
USER_2	18-23	Uni. entrance qualification	No	Intermediate	Written
USER_3	23-27	Master	Yes	Beginner	Written
USER_4	23-27	Bachelor	Yes	Advanced	In person
USER_5	43-47	Master	No	Advanced	Written
USER_6	23-27	Bachelor	Yes	Advanced	Written
USER_7	28-32	Staatsexamen (State Examination)	No	Beginner	In person
USER_8	33-37	Master	No	Advanced	In person

6.2.3 *New Usability Problems*

During the second iteration, several interface issues were identified and are summarized below.

- **UP-12**: Visual framing and placeholder names in the onboarding flow caused participants to perceive the screen as a static tutorial, discouraging configuration changes (USER_3).
- **UP-13**: A numerical indicator was requested instead of a single blue checkmark to more accurately reflect the number of mutual connections, since verification by multiple independent users provides stronger security guarantees than a single connection (USER_5).
- **UP-14**: Participants recommended that the interface explicitly state that the feature applies only to in-app contacts to prevent users from assuming that the messenger accesses their entire phone contact list (K32BL).

6.2.4 *Discussion*

This iteration demonstrated a significant improvement over the first. Alongside the directly identified usability issues, a substantial part of the PUW involved discussions

about the features and actively performing the verification. Below are the key insights from these discussions which are enriched with the in-app statistics from Figure 9 and Figure 10, alongside the feedback provided by external users in Table 4.

1. **Increase trust through transparency**: Subtle details, like displaying the verification date, or the option for manual approval, did enhance the perceived trust in the messenger. These findings also aligned with the results of Kebbedies [16].

2. **Non-technical users tend to use default settings without reading**: The data from the in-app study (Figure 9) correlated with the face-to-face feedback and the discussions during the PUW, showing that most participants retained the default settings and only technical users chose alternative configurations, such as enabling manual approval or disabling transitive trust sharing based on personal preferences.

3. **Active scanning and missing threat model**: A significant portion of the discussion focused on whether participants would actively scan a contact's QR code. Non-technical users primarily wanted to eliminate the red verification badge because they found it annoying. Most participants abstained from in-person verification due to convenience, reflecting a general absence of a perceived threat model. Even professional participants who audit cryptographic protocols in their daily work did not actively verify safety numbers, convinced that the platform should manage verification automatically. This behavior aligns with the diary-based findings of Fassl [17]. It is also supported by the in-app statistics in Figure 10, where the median number of verified contacts remained constant between weeks W15 and W17. However, after introducing the WoTM feature in week W18, both the average and median values increased.

6.3 SUMMARY

The results of the PUWs and the additional feedback indicated that most users had no objections to the contact-sharing mechanism proposed by the WoTM, which restricts visibility to mutual friends. The verification rate increased substantially after the introduction of WoTM, whereas the interface modifications designed to simplify manual verification produced no comparable increase. The in-app statistics thus matched the statements of several participants who would not verify manually even when presented with a red warning badge. Nevertheless, some participants stated that details such as the verification badge and subtler transparency features like the verification date could increase trust in the messenger.

Most in-app study participants left the WoTM feature enabled, with only a few technical users opting out. Non-technical users generally retained the default settings and often skipped the onboarding texts, while only technical users chose alternative configurations such as enabling manual approval. These findings motivated a two-path onboarding design: a default path with predefined settings and a customized path that allows users to configure the WoTM manually.

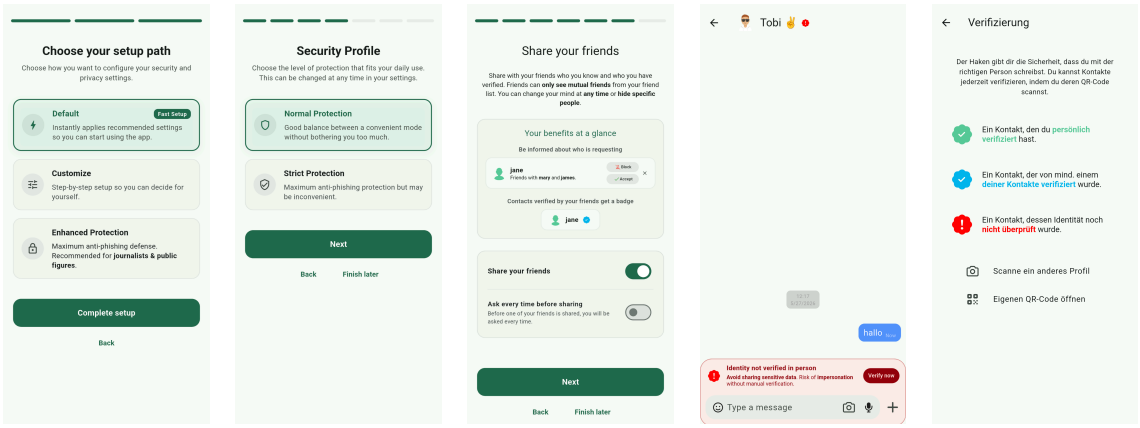
The SUT iterations in the following chapter therefore evaluate whether users accept this design and whether the modifications in Section 5 help them complete the ceremony.

7 RESULTS OF THE SIMPLIFIED USER TESTING

This chapter presents the results gathered during the SUT as described in Section 4.2.2. The goal of the SUT was to evaluate how users respond to the setup path that resulted from the PUW iterations described in Section 6. Additionally, the modifications to the authentication ceremony proposed in Section 5 were tested to determine whether they enable users to perform the authentication ceremony more easily. For this purpose, three iterations were conducted, each consisting of five participants. During a session, participants were asked to perform the tasks described in Section C. For each iteration, the modifications performed, the quantitative data collected along with the respective improvements, and the qualitative data gathered in the form of usability problems the participants encountered are described. Each iteration concludes with a short discussion of the key insights from the respective iteration.

7.1 ITERATION 3 - FIRST SIMPLIFIED USER TESTING

The primary modification for this iteration was the introduction of the "setup path" (a) during onboarding. This allows users to choose between default and customized configurations. Additionally, security profiles (b) were introduced to address the needs of journalists or public figures who have a different threat model than non-technical users, for whom such a warning could reduce trust in the messenger. When enabled, this strict mode displays a red warning message for unverified chats (d) and makes the red verification badge more prominent throughout the application. Furthermore, clicking on the verification badge or initiating contact verification displays the explanation screen shown in (e), describing the purpose and usage of the badges.



(a) (b) (c) (d) (e)

Figure 11: This figure shows the main changes that were tested in this iteration.

7.1.1 *Quantitative Results*

Table 5: The left table shows a demographic summary of the participants of iteration 3. Participants who completed the task successfully are marked in green. The right table shows the mean and standard deviation for the quantitative data.

ID	Age	Highest Degree	Student	ATI	Security Knowledge
J56M3	53-57	Diplom	No	2.89	Intermediate
AN532	53-57	Diplom	No	3.11	Beginner
AH89I	28-32	Bachelor	Yes	5.11	Beginner
M37AB	23-27	Bachelor	Yes	5.22	Advanced
J1M53	18-22	Bachelor	Yes	5.0	Advanced

Metric	**Value**
Pass rate	3/5 (60%)

Metric	Mean	σ
Completion time	1m 34.67s	1m 12.78s
SEQ	4.2	1.1
UX	3.8	0.84
UMUX-Lite	4.6	0.55
Trust	4.6	0.55
Understanding	3.8	1.1
Recommend	8.2	0.84

7.1.2 *New Usability Problems*

This section lists the main usability problems that occurred in the third iteration.

- **UP-15**: The term "verify" was unfamiliar to participant AN532 due to a lack of social media experience. This led them to enable the Strict Security Profile instead of clicking the "Verify user" button in the general settings, subsequently ignoring the resulting red warning.
- **UP-16**: Participants AH89I and AN532 did not understand the purpose of the red verification badge because both chose the default setup path, which did not include the verification badge explanation. Participant AN532 was only able to verify the other user after completing the third task, which explained the verification badges. Participant AH89I understood the purpose after reading the warning of the strict security profile (d). However, due to UP-18, AH89I was still unable to complete the verification.
- **UP-17**: The in-app notification displayed after a successful verification was unclear. This left participant J56M5 unsure whether the contact was verified, which resulted in a high task completion time.
- **UP-18**: The verification badge example page (e) was perceived as interactive because participant AH89I assumed the badges were clickable. Additionally, the label "scan

other profile" was misunderstood as referring to another person rather than the contact's QR code, preventing AH89I from initiating the verification flow.

7.1.3 *Discussion*

Using the qualitative feedback and discussions from the post-test interviews, four key insights emerged from this iteration:

1. **The setup path worked as intended**: Participants with security experience (M37AB, J1M53) chose the customized path, whereas the remaining participants opted for the default path and, after learning about its features during task 3, considered the default configuration appropriate and enabled the WoTM feature.

2. **Wrong explanation placement**: The verification badge explanation should precede the setup path choice because participants who chose the default path did not understand the red badge initially.

3. **Security Profiles could create a false sense of security**: Participant AN532 failed to solve the task, as they mistakenly believed that enabling maximum protection in the security profile settings was sufficient. Similarly, participant AH89I also enabled this setting and, after failing to understand the verification badge example page UP-18, assumed that maximum protection was enough.

4. **Impersonation prevention and gamification as stronger motivators than a theoretical MITM attack**:

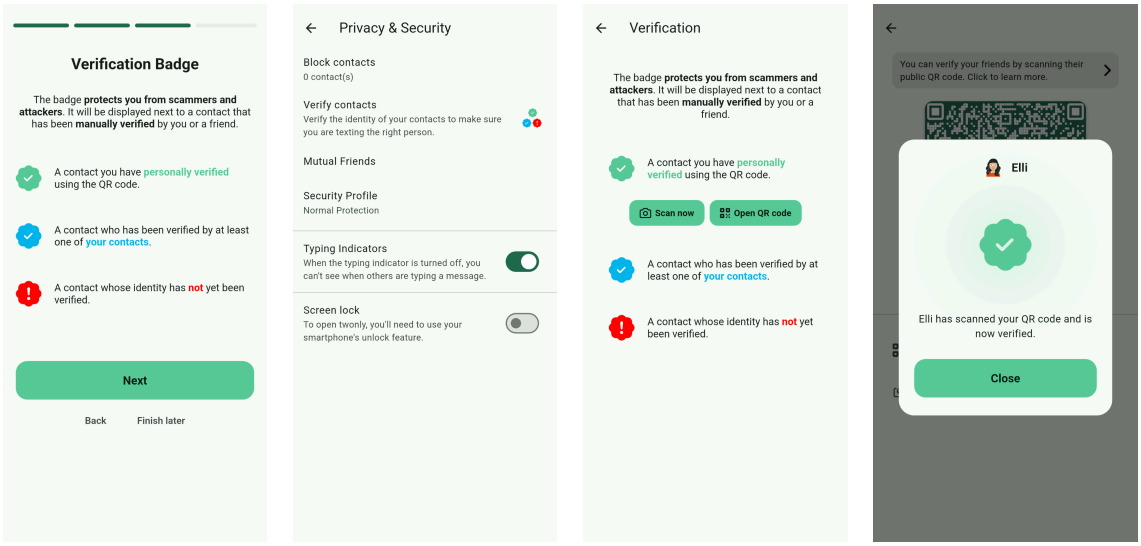
   - Only the two participants with advanced security knowledge understood the cryptographic rationale behind verification, whereas the remaining participants did not. Even after an explanation by the moderator, they did not fully comprehend this reasoning. In contrast, the risk of impersonation was understood by all participants, serving as a stronger motivator. For instance, participant AH89I explained that impersonation attempts were common in their corporate environment, and that verification could help prevent successful attacks.

   - Furthermore, political risks can prevent trust in messengers entirely, even if the applications claim to be secure. A participant who was born in Turkey (AH89I) refused to trust secure messengers because individuals faced imprisonment there simply for installing secure applications like ByLock [44]. Consequently, they restricted their usage to WhatsApp for non-sensitive communication due to concerns that communication partners could share screenshots with the authorities.

   - Participant M37AB mentioned that their motivation to verify contacts stemmed from a desire to remove the red badge and that they enjoyed verifying others, suggesting that introducing gamified rewards, such as custom animations or additional features, could further enhance user engagement.

7.2 ITERATION 4 - SECOND SIMPLIFIED USER TESTING

To address the problems identified in the third iteration, the fourth iteration focused on improving the verification badge explanation. The main change was moving the badge explanation into the onboarding setup (a), so that users could learn about the badges before encountering them in the application. The verification badge screen was also updated by placing the action buttons directly below the green badge. Additionally, a short subtitle about the badges was added to the system settings (b). Finally, after a successful scan, a clear confirmation message was shown to both participants (d) to reduce the uncertainty about whether the verification had completed successfully.



(a) (b) (c) (d)

Figure 12: This figure shows the main changes that were tested in this iteration.

7.2.1 *New Usability Problems*

This section lists the main usability problems that occurred in the fourth iteration.

- **UP-19**: AD569 and H98HL did not understand the verification badge explanation shown in Figure 12 (c). Both participants attempted to verify themselves against the provider instead of scanning the moderator's QR code. AD569 expected to scan a printed QR code which he did not find, while H98HL searched for an option to verify by scanning their face.
- **UP-20**: All three participants who were able to solve the task were unsure, while seeing the success popup, whether they had already verified successfully. They were only sure after going back to the contact view, where the contact was listed.
- **UP-21**: The blue verification badge shown during onboarding confused most participants and was not correctly understood during the post-test interview questions.

- **UP-22**: The verification badge explanation during onboarding was either not read by the participants or misunderstood as solely an informational page, while participants expected to actively perform some actions.

7.2.2 *Quantitative Results*

Table 6: The left table shows a demographic summary of the participants of iteration 4. Participants who completed the task successfully are marked in green. The right table shows the mean, standard deviation, and percentage improvements relative to the previous iteration for the quantitative data.

ID	Age	Highest Degree	Student	ATI	Security Knowledge
AD569	43-47	Staatsexamen	No	2.44	Intermediate
LB590	28-32	Master	No	1.56	Intermediate
EB389	23-27	Master	No	1.78	Beginner
L568B	28-32	Master	No	5.11	Intermediate
H98HL	23-27	Master	Yes	2.44	Beginner

Metric	**Value**	**Improvement**
Pass rate	3/5 (60%)	+0%

Metric	Mean	σ	Improvement
Completion time	1m 27s	47.66s	+8%
SEQ	4.4	0.55	+14%
UX	4.4	0.55	+44%
UMUX-Lite	4.8	0.45	+21%
Trust	4.6	0.55	+0%
Understanding	4.2	0.45	+26%
Recommend	6.2	3.49	−36%

7.2.3 *Discussion*

Using the qualitative feedback and discussions from the post-test interviews, the following key insights emerged from this iteration:

1. **Misconception of the verification target:**
   - AD569 directly spotted the red verification badge and correctly identified that the contact might be fake. However, the participant did not use the in-app verification flow. Instead, the participant attempted to verify the contact by exchanging personal questions and an image, and then declared the contact successfully verified even though the red badge was still displayed.
   - H98HL followed the same approach. When both participants were asked to click the verification badge, neither understood how to proceed (see UP-19), as

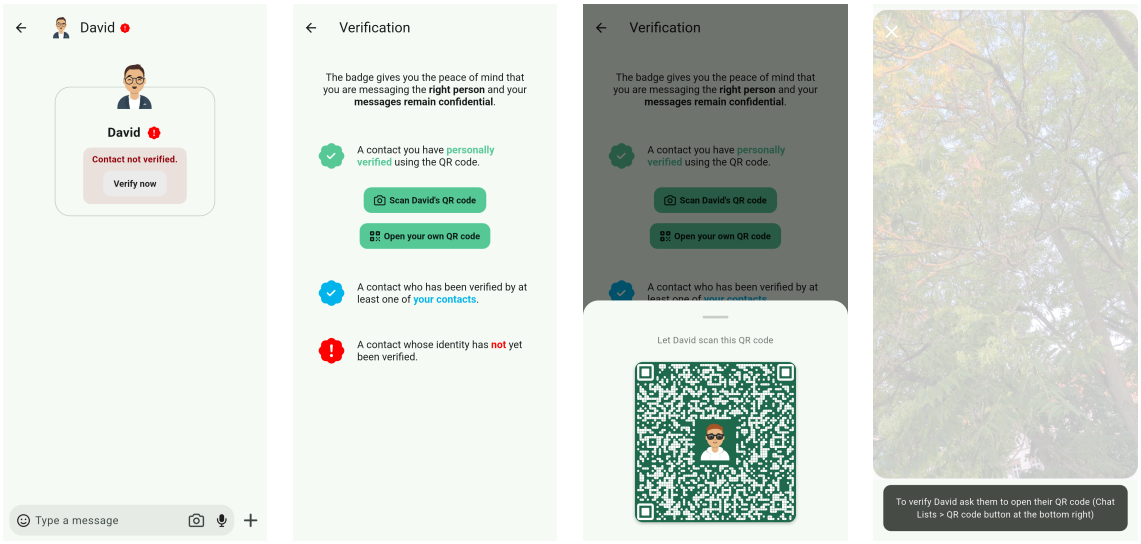
both expected to authenticate themselves against the platform rather than scan the other user's QR code.

- Ultimately, both failed to solve the task, as they did not understand which QR code they had to scan and believed that verifying through the context of the chat would be sufficient.

2. **The setup path again worked as intended:** As in the previous iteration, only the technically experienced participant chose the customized setup path, while the remaining participants followed the default configuration. During task three, all participants who selected the default branch enabled the WoTM feature as well as the manual approval setting. In the post-test interview, participants stated that they were comfortable with the manual approval not being the default configuration, as they could change it later.

3. **Explanations are most effective when provided in context:** The blue verification badge shown during the onboarding setup confused most participants, as its meaning was not clear without prior context (e.g. sharing their friends). Also, as demonstrated by H98HL, the onboarding slides were skipped without being read.

7.3 ITERATION 5 - THIRD SIMPLIFIED USER TESTING



(a) (b) (c) (d)

Figure 13: This figure shows the main changes that were tested in this iteration.

The main change in the fifth iteration was adding a contact status overview to the chat (a), similar to Signal's phishing prevention [45], but with a "Verify now" button to directly open the verification page (b). Additionally, the onboarding explanation slides were removed, and text across the application was updated to clarify that users must scan the other user's QR code. Specifically, descriptive messages were added next to the

QR code (c) and in the camera view (d). After successful verification, the application navigated back to the contact view showing the green verification badge. Finally, the security profiles introduced in the third iteration were also removed, as they had caused more confusion than they resolved.

7.3.1 *Quantitative Results*

Table 7: The left table shows a demographic summary of the participants of iteration 5. Participants who completed the task successfully are marked in green. The right table shows the mean, standard deviation, and percentage improvements relative to the previous iteration for the quantitative data.

ID	Age	Highest Degree	Student	ATI	Security Knowledge
AD2PK	23-27	Uni. entrance qualification	Yes	3.89	Intermediate
TM972	23-27	Bachelor	Yes	3.78	Advanced
LU896	23-27	Apprenticeship	Yes	3.33	Intermediate
F9KLP	23-27	Uni. entrance qualification	Yes	2.33	Beginner
KUV87	33-37	Bachelor	No	4.78	Advanced

Metric	Value	Improvement
Pass rate	5/5 (100%)	+67%

Metric	Mean	σ	Improvement
Completion time	1m 22.6s	27.84s	+5%
SEQ	4.4	0.55	+0%
UX	4.6	0.55	+16%
UMUX-Lite	4.8	0.45	+0%
Trust	5	0	+90%
Understanding	4.4	0.55	+14%
Recommend	8.4	2.07	+66%

7.3.2 *Recurring Usability Problems*

During iteration 5, only one usability problem occurred, which is related to UP-20:

- **UP-23**: Participants AD2PK and LU896 were not immediately confident that the verification was successful. Because both participants opened their QR codes instead of using the camera, the verification screen did not close automatically. Similar to UP-20, this resulted in both users navigating back to the contact view to confirm that the red verification badge had disappeared.

7.3.3 *Discussion*

Using the qualitative feedback and discussions from the post-test interviews, the following key insights emerged from this iteration:

- **Signal's contact overview page proved effective:** All participants utilized the verify button within this interface and correctly inferred from the prompt that the conversation might lack protection.
- **The setup path worked again as intended:**
  - Participant AD2PK selected the customized settings because they always preferred maximum-privacy configurations but considered the feature's default activation understandable and helpful for most users.
  - Participant LU896 used the default setup path and later enabled manual approval. During the interview, the participant stated that it was understandable for the feature to be enabled by default during setup.

7.4 SUMMARY

The SUT refined the authentication ceremony over three iterations, culminating in a 100% task success rate in the final iteration. No usability issues were identified in the final iteration that prevented participants from completing the authentication ceremony. However, two participants still experienced uncertainty about successful completion (UP-23). The key insight from the three iterations is that users require a clear, highly guided path through the UI, featuring explanations provided precisely at the moment they are needed, using straightforward instructions that leave no room for misinterpretation.

In earlier iterations, task failures occurred primarily because users either developed a false sense of security or misunderstood the process entirely. Some users believed that strict privacy settings or in-chat questions were sufficient, while others attempted to authenticate themselves against the service provider instead of scanning their contact's QR code. Contextual explanations and a clearer UI flow in the final iteration successfully resolved these misconceptions.

The setup path introduced in the first SUT iteration (Section 7.1) proved highly effective by showing the WoTM settings only if the user opted for a customized configuration. Non-technical users consistently chose the default path, while technical users opted for the customized path. Ultimately, both groups were satisfied with their respective configurations, as they either appreciated the default features or were able to easily disable them according to their personal privacy preferences.

8 RESULTS OF THE SURVEY AND IN-APP USAGE

This chapter reports the comparative online survey (Section 4.5) and in-app usage statistics (Section 4.3) for the redesigned ceremony. These results complement the PUW findings in Section 6 and the iterative SUT results in Section 7 with users' preferences and observations of everyday verification.

8.1 SURVEY DEMOGRAPHICS AND FEATURE AWARENESS

After the final modifications of the SUT were released for multiple weeks, the comparative online survey was published and promoted on Mastodon [37] with the questions listed in Section E. In total, 23 users participated in the survey with a median ATI score of 4.56, indicating a technically inclined sample. More respondents reported awareness and prior use of the ceremony in twonly than in Signal (Table 8). The 52.2% who reported having verified contacts in Signal reflect this technically inclined sample rather than Signal users in general (Section 10.5).

Table 8: Comparison of verification feature awareness and usage.

	Signal	**twonly**
Median Usage Duration	> 2 years	1 year
Authentication Ceremony Awareness	87.0%	100.0%
Authentication Ceremony Performed	52.2%	82.6%

8.2 REASONS FOR VERIFICATION AND NON-VERIFICATION

In Signal, the main reasons for not verifying were a lack of knowledge and laziness, while laziness also applied to twonly. Multiple users reported that they saw no reason to verify, as they had already confirmed the identity of their contacts through the context of the chat or via another channel. For twonly, multiple participants noted they only wanted to try out the feature, not because of the verification itself.

These statements are supported by the in-app statistics shown in Figure 14. Following the introduction of the verification badge, the QR code verification rate increased only slightly and remained mostly stable over time. This occurred even though the badge in twonly is more visible and no usability issues were identified in the final iteration (Section 7.3) that prevented participants from completing the authentication ceremony.

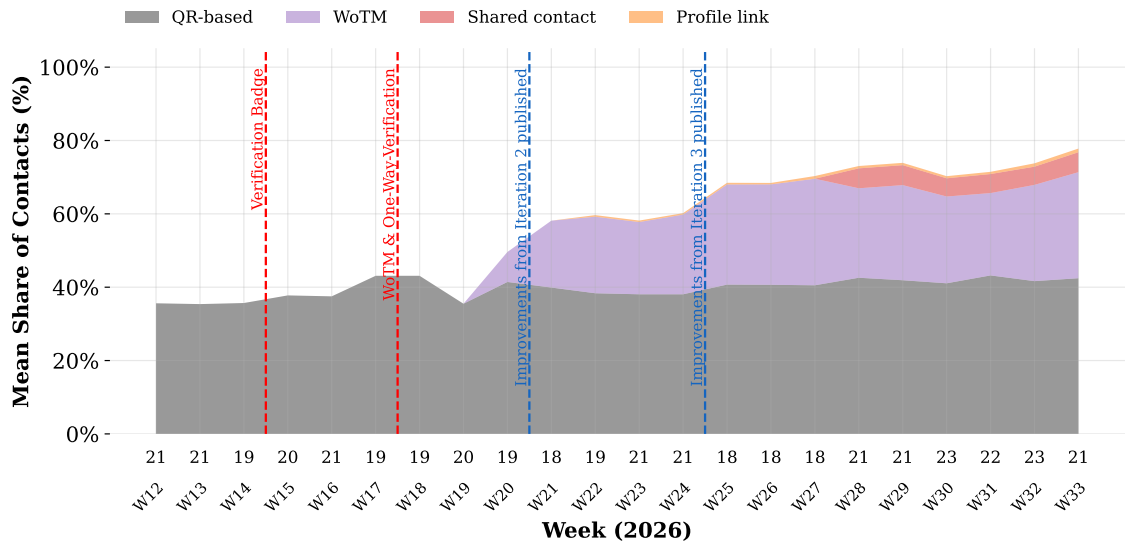


Figure 14: This figure shows the weekly mean participant-level share of contacts by verification method. QR-based verification combines direct scans and authenticated notifications after a scan (Section 5.3.2 and Figure 1). The other categories are profile-link verification (Section 5.3.4), shared-contact verification (Section 5.4.1), and automatic WoTM propagation (Section 5.4). The counts above the week labels show how many participants opened the application and supplied data in that week, rather than cumulative enrollment (see Section 4.3). Only the latest record per participant and week was retained, and values from earlier weeks were not carried forward. The count was 21 in the final week. In that week, participants had verified a mean of nearly 80% of their contacts, including about 40% through QR-based verification and 28.9% through WoTM.

The 35% initial verification rate at the start of the data collection in Figure 14 is consistent with feedback from multiple survey participants, who reported that most verifications occur when adding contacts via QR code. In twonly, this process is prominent and automatically verifies both parties. Signal also supports contact initiation through username QR codes, but this flow remains separate from identity-key verification [46]. The relevant difference is therefore the integration of contact exchange and verification.

8.3 AUTOMATED VERIFICATION AND TRANSITIVE TRUST

The in-app statistics in Figure 9 show that the majority of in-app study participants enabled the automated verification feature, consistent with the choices observed during SUT onboarding. The introduction of WoTM coincided with the largest observed increase in the overall verification rate. As shown in Figure 14, WoTM accounted for a mean participant-level share of 28.9% of contacts in the final week.

8.4 USABILITY PRIORITIES AND PREFERENCE

Respondents rated how important six aspects were to them when performing contact verification, from 1 (not important at all) to 5 (very important) (Section E.2.1). These

ratings describe priorities, rather than satisfaction with either messenger. Simplicity, visual feedback, and trust in the security of the process received median ratings of 5.0. Simplicity received the most maximum ratings, with 19 of 23 respondents selecting 5 (82.6%). Process duration, comprehensibility, and discoverability had median ratings of 4.0 (Table 9).

Table 9: Importance of six aspects when performing contact verification, rated by 23 comparative online survey respondents from 1 (not important at all) to 5 (very important).

Aspect	**Median**	**Rated 5, n (%)**
Duration of the process	4.0	10 (43.5%)
Simplicity of the procedure	5.0	19 (82.6%)
Comprehensibility of the workflow	4.0	11 (47.8%)
Visual feedback	5.0	14 (60.9%)
Trust in the security of the process	5.0	16 (69.6%)
Discoverability of the verification feature	4.0	9 (39.1%)

In the related free-text question about motivation and important aspects, simplicity was also the most frequently cited requirement (Section E.2). A separate question asked which verification process respondents preferred for everyday use: 65.2% chose twonly, 17.4% chose Signal, and 17.4% preferred neither version.

8.5 PERCEIVED SECURITY

For perceived security, 34.8% reported no difference, 34.8% favored twonly, and 30.4% favored Signal. Respondents cited Signal's reputation as a secure messaging standard praised by security experts, or twonly's project transparency and personal acquaintance with its developer, as reasons for their assessments.

8.6 TERMINOLOGY

Over half of the participants (56.5%) found the term "verification badge" clearer than "safety number". Another 30.4% saw no difference, and 13.0% preferred Signal's terminology. Participants generally considered verification badges more familiar due to their usage on other social media platforms and viewed the term "safety number" as non-explanatory. Only one technical user opposed the new terminology, arguing that a verification badge is a mere interface artifact, whereas "safety number" correctly reflects the underlying cryptographic process.

8.7 VERIFICATION BADGES

The verification badges introduced a prominent interface change. Although all participants were aware of the authentication ceremony, only 87.0% noticed the new verification badges. Despite their visibility, 73.9% reported no change in their overall messenger usage. The red warning badge motivated only 39.1% of the users to verify their contacts, while another 39.1% ignored it.

Despite this reluctance, 18 of 23 participants (78.3%) stated that they would like verification badges to be implemented in Signal.

8.8 SUMMARY

Most survey respondents preferred twonly's verification process and found "verification badge" clearer than "safety number". Simplicity, visual feedback, and trust in the process received the highest median importance ratings. In the usage data, manual verification changed little, while the introduction of WoTM coincided with the largest increase in verified-contact coverage. In the final week, WoTM accounted for a mean participant-level share of 28.9% of contacts among the 21 participants with data.

9 DISCUSSION

The main contributions are implementation and deployment evidence for previously proposed designs, together with observations of their limits. This chapter first discusses these findings, then summarizes the design recommendations they confirm.

9.1 AUTOMATE TRUST TRANSPARENTLY

Prior work had already proposed trust through friends and distinct visual trust states [15, 27]. The WoTM added an implementation and observations from its use in a deployed messenger. In the final week, it accounted for a mean participant-level share of 28.9% of contacts among the 21 participants with available data (Figure 14). Its introduction coincided with the largest observed increase in verified-contact coverage, while direct QR verification changed little.

These findings support automation as a way to extend verification beyond manually verified contacts, following Dykstra's paradigm of invisible security, in which defenses run without user intervention [47].

9.2 REDUCE THE STEPS REQUIRED FOR MUTUAL VERIFICATION

Kebbedies [16] recommended reducing the number of steps required for mutual verification. This thesis implemented this recommendation through a common profile QR code and an authenticated notification that allows both contacts to be verified after one scan (Section 5.3.1). The comparative survey confirms that users prioritize this simplicity: it received a median importance rating of 5.0 and the highest number of maximum ratings among all six aspects (19 of 23 respondents, 82.6%), and it was the requirement named most often in the free-text answers (Table 9). These results support the broader recommendation that security tasks should be practical and time-saving [48].

9.3 CLARIFY WHO VERIFIES WHOM

The terminology results qualify the recommendation to replace unfamiliar cryptographic terms [16]. In the comparative survey, 56.5% found "verification badge" clearer than "safety number" (Section 8). However, two participants in an earlier iteration expected to verify themselves against the provider (UP-19). Familiarity with platform-issued badges did not ensure understanding of contact verification. The term and explanation were evaluated together, so the misunderstanding cannot be attributed to the term alone.

The implication is to explain who performs the check and whose identity is being checked, alongside the familiar label. The final iteration made the instruction to scan the other person's QR code explicit (Section 7.3). Terminology should therefore be assessed through users' actions and explanations as well as their stated preference.

9.4 ADDRESS MOTIVATION BEYOND USABILITY

The survey and in-app results (Section 8) showed that a usable ceremony does not guarantee adoption. Fassl et al. [17] found that social friction and a lack of perceived benefit block adoption even when the design is usable. Our data shows these same barriers remain: all five participants completed the final ceremony (Section 7.3), yet manual verification barely increased after the redesign was published (Figure 14).

The ceremony prevents impersonation, above all by the server that distributes the keys. Kebbedies found that this server is not part of most users' threat model, because they trust the provider to keep their chats secure [16]. This thesis therefore framed the ceremony not only as protection against the server, but also as a way to confirm the identity of the other person and prevent impersonation by other users (Section 5.2). The iterations and the survey show that this framing did not replace how participants established identity. As with the server, most of them trusted the conversation itself and confirmed a contact through the messages exchanged or through a second channel, rather than through the ceremony (Section 7.1, Section 8). This helps explain why manual verification did not increase after the redesign, and it leaves open which further benefit would motivate users to verify (Section 11.1).

9.5 CONTEXTUAL AND REPUTATION-BASED TRUST

Trust in secure messaging proved to be largely contextual and reputation-driven, independent of the authentication ceremony itself. The perceived security data in Section 8 reveals that participants justified their trust in Signal through its institutional reputation, and their trust in twonly through personal acquaintance with the developer, not through any evaluation of the cryptographic properties of either system. This has a significant implication for authentication ceremony design: improving the ceremony may be necessary but not sufficient for better security outcomes, because users actually form trust through social rather than technical mechanisms. In extreme cases, this relationship breaks down entirely. Participant AH89I from Turkey reported avoiding secure messengers altogether because installing applications such as ByLock had led to imprisonment in their home country (Section 7.1). This illustrates that political threat models can render authentication ceremony design irrelevant, pointing to a dimension of the problem that is outside the reach of interface design.

9.6 PROVIDE INSTRUCTIONS AT THE POINT OF ACTION

The iterations confirmed earlier recommendations for discoverable controls and clear feedback [16, 26]. Onboarding explanations were skipped or misunderstood (UP-22). In the final iteration, the contact overview displayed the verification state with a "Verify now" button, placing the action directly in the chat (Figure 13). All five participants used it. Placing the prompt in the chat therefore made the action discoverable without relying on onboarding text.

9.7 PROVIDE SENSIBLE DEFAULTS

A consistent finding across both the PUW and all SUT iterations is that non-technical users adhered to the default settings, whereas technical users actively customised their configurations. The two-path onboarding design proved successful because users who selected the default setup were not surprised by the features that were automatically enabled. Even participants who later adopted more restrictive privacy settings considered the default choices appropriate for a general audience. They were comfortable choosing the default path during setup because they expected the defaults to be reasonable, knowing they could adjust or disable specific features later in the application settings.

10 LIMITATIONS

This chapter acknowledges the limitations of the methodology and evaluation, discussing concrete threats to validity identified during the research.

10.1 IMPLEMENTATION IN TWONLY AND TRANSFERABILITY TO SIGNAL

The proposed modifications were implemented and evaluated in twonly rather than Signal, and they differ in how far they could be transferred. The redesign combined the following core elements: the changed terminology with its verification badges, the one-way QR code scan that verifies both contacts, and the automated WoTM. While the first element requires only changes to the user interface, the other two require more complex changes to how contacts and keys are exchanged.

The one-way scan depends on an in-person contact exchange that can carry the identity key. Signal already offers a QR code as part of its username feature, which lets users be contacted without sharing their phone number [46]. This code resolves a username, but it currently does not authenticate the contact's identity key. An in-person variant could additionally contain the public identity key, allowing the scanning device to compare it with the key used by the resulting session. Such a change must account for identity keys that change, for example after reinstallation or replacement of the primary device [10], which would leave a previously shared code stale. This is not a major obstacle, as twonly's verification logic likewise limits the lifetime of a QR code to one day (Figure 1). Signal could also generate two codes: the existing resettable username link for sharing publicly, and a short-lived code containing the identity key for in-person contact exchange. Such a code would verify the scanned contact only. The authenticated notification in Section 5.3.3, which verifies both contacts after a single scan, could be implemented in Signal as well, but it would require a new message to be exchanged between the clients.

The WoTM would be the hardest element to transfer. The mechanism requires a user to allow their verified contacts to learn which other contacts they have verified. Signal currently keeps a user's contact list private, so this would fundamentally change what Signal reveals about a user to their contacts [46]. A core value of Signal is that its server learns almost nothing about a user, as its private contact discovery for phone numbers illustrates [49]. The porposed WoTM would preserve this property, because only mutual contacts learn about a verification and the social connection it implies, while the server learns neither. A practical path would follow the privacy controls Signal already offers alongside usernames, where users decide who can reach them by phone number and who can see it [46]. Sharing verified contacts could be added as a further setting of this kind, one that governs information beyond the phone number, namely the verification state of a user's contacts. The PUW results support this direction, as the majority of participants raised no objections during the discussion of this feature to such a mechanism in Signal, whose visibility is restricted to mutual contacts.

10.2 SECURITY VALIDATION AND TRUST ASSUMPTIONS

The one-way verification and WoTM protocols proposed in this thesis were neither independently audited nor formally validated. The thesis instead focused on the usability of the redesigned authentication ceremony and whether users would accept the proposed mechanisms. The evaluation therefore supports conclusions about how participants understood, used, and perceived the designs, but it does not establish their cryptographic security.

The study also did not systematically investigate how long WoTM verification information should remain valid or how transitive trust should respond when a friend's device is compromised or multiple friends collude. These protocol properties and trust assumptions require dedicated security analysis and independent review.

10.3 EXPLORATORY SCOPE AND SAMPLE SIZE

A total of 21 participants took part in the usability studies, 32 in the in-app study, and 23 in the comparative online survey. The usability sessions were designed to find problems, not to measure how often they occur. Five participants per iteration reveal most of the problems in an interface, as Nielsen recommends (Section 2.2), and each iteration used its findings to revise the design. These sessions therefore show which problems existed and whether a revision removed them. That all five participants completed the ceremony in the final iteration means that no blocking problem remained in these sessions, not that the ceremony works for every user.

Two further constraints limit what the numbers can show. Each iteration changed several design elements at once and involved different participants, so an improvement cannot be traced back to a single change. The quantitative results were therefore treated descriptively, and Nielsen's arcsine transformation only compares the iterations with one another instead of estimating success rates for a wider population.

The number of participants represented in the in-app data varied between weeks. If more than one record was available for a participant in a given week, only the latest record was included. In the final week, data from 21 participants were available. Since the weekly means were calculated from different groups of participants, some of the observed differences may be related to changes in participant composition.

10.4 PARTICIPANT BIAS IN IN-APP AND COMPARATIVE ONLINE SURVEYS

Representativeness limits generalization independently of sample size. Both surveys recruited existing twonly users, many of whom personally knew the developer (the author of this thesis). This may have favored positive assessments of twonly, despite requests for honest answers and critical feedback. Increasing the sample through the same recruitment channels would not necessarily resolve this limitation.

10.5 CONVENIENCE SAMPLING AND ATI SKEW

The in-app and comparative online surveys, both promoted on Mastodon, showed a skew towards technically experienced users. In the in-app sample, approximately 62% reported advanced computer knowledge and 38% reported advanced IT-security knowledge. In the comparative online sample, the corresponding proportions were 70% and 65%, and the median ATI score was 4.56. Consistent with this skew, 52.2% of that sample reported having verified contacts in Signal, so their preferences cannot be assumed to represent users unfamiliar with verification. The SUT reduced this bias by pre-screening participants for a balanced sample, but the comparative online and in-app surveys lacked those controls. As a result, the highly positive reception of the WoTM and verification badges might not apply to less technically experienced users.

10.6 LANGUAGE AND CULTURAL BIAS

All study sessions, interviews, and materials were in German. Whether the modifications would yield similar results in other languages and cultures remains an open question. Terminology effects may differ across languages, because mapping technical concepts like “safety number” to more accessible terms relies on culturally specific conventions that might not translate directly.

11 FUTURE WORK

Building on the findings and identified limitations, this chapter outlines concrete directions for future research.

11.1 MOTIVATION, INCENTIVES, AND GAMIFICATION

The motivational barriers discussed in Section 9.4 call for evaluating benefits users receive from verification. Participant M37AB suggested gamified rewards, and comparative survey respondents suggested custom animations or access to features after verifying a contact. Their effects on verification behavior were not evaluated. Future studies should measure voluntary verification in everyday use as well as task completion and stated preference.

The passwordless backup mechanism introduced by twonly [50] provides a concrete candidate: users can select only verified contacts to hold private key shares. This offers a practical reason to verify and may reduce the social friction identified by Fassl and Krombholz [17]. The feature was released after in-app data collection ended, so it did not affect the reported observation period. Whether it increases manual verification remains to be tested.

11.2 EVALUATE EFFECTIVENESS AGAINST PHISHING

This thesis focused primarily on the usability of the authentication ceremony itself rather than on attack simulation. Whether the verification badge actually prevents phishing in practice was not empirically tested. Introducing security profiles in the third SUT iteration also created a false sense of security for some users, who believed enabling maximum protection was enough without actually performing the verification ceremony. Future adversarial user studies, such as simulated phishing attacks against verified and unverified contacts, will need to evaluate whether these UI elements mitigate threats or just create overconfidence.

11.3 BLUE VERIFICATION BADGE COMPLEXITY

Some users found the concept of the blue verification badge (representing transitive trust) too complex. A notable issue was color confusion: contrary to the intended hierarchy, where direct in-person verification (green badge) represents the highest level of trust, users associated the blue badge with a higher level of trust due to social media conventions, such as verified accounts on Twitter or YouTube. Future user studies should focus on finding an optimal visual encoding that clearly distinguishes direct trust (from QR scans) from transitive trust (from the WoTM).

11.4 TRUST DECAY AND RE-VERIFICATION

Currently the trust established by the WoTM has no expiry. This raises open questions regarding trust decay: if Bob verifies Alice and Charlie, but Bob's device is later compromised, Alice's and Charlie's blue badges become potentially invalid. Future research must determine whether transitive trust should require periodic re-verification and how to balance security requirements with user burden.

11.5 INTEGRATION INTO ESTABLISHED MESSENGERS

Section 10.1 discusses which parts of the redesign could be transferred to Signal and what each of them would require. Future work should evaluate these steps in practice. The interface changes, such as the revised terminology, the verification badges, and the instructions shown at the point of action, could be tested in Signal on their own. The in-person QR code and the authenticated notification would first have to be implemented, and the WoTM would additionally require a decision on how verified contacts are shared.

12 CONCLUSION

Building on previous research by Kebbedies and Fassl, we contributed a redesigned authentication ceremony with simpler manual verification and automated verification via the proposed Web of Trust for Messengers (WoTM), its implementation in an open-source messenger, and its evaluation with real users. Following a user-centered methodology, two Pluralistic Usability Walkthroughs (PUWs) with six participants evaluated an interactive UI prototype and screenshots of the resulting implementation, and informed the redesign in twonly [30]. Its public release enabled the collection of real-world usage data beyond the laboratory. We then conducted an iterative Simplified User Testing (SUT) with 15 participants to polish the implementation and fix newly introduced usability issues, achieving a 100% success rate in the final iteration. Finally, we evaluated the ceremony with in-app usage statistics from 32 users and a comparative online survey of 23 twonly and Signal users.

The deployment added evidence that the laboratory could not provide. Most survey respondents preferred the simplified ceremony in twonly and found "verification badge" clearer. In the in-app data, the WoTM produced the largest observed change in real-world usage, automatically covering a mean of 28.9% of contacts in the final week, whereas manual verification barely increased, suggesting that motivational barriers remained despite the usability improvements.

To conclude, we revisit the research question defined for this work and summarize how it was answered:

- RQ: *How could Signal's authentication ceremony be improved to be more user-friendly?*
  - **Provide an intuitive, step-by-step flow:** Guide users along a clear, linear path with short instructions shown exactly where they act. Replace cryptographic jargon with everyday language that makes clear who verifies whom, and confirm success immediately and visibly.
  - **Prioritize simplicity and speed:** Using a single, central QR code for one-way verification minimizes user effort. Furthermore, complex settings during the setup should be hidden behind sensible defaults to accommodate non-technical users, while still allowing technical users to customize them.
  - **Automate trust transparently:** Integrate verification into interactions such as scanning a QR code to add a contact. Contact-based trust propagation can extend verified-contact coverage, but its status and reliance on trusted contacts should remain distinguishable from direct verification.
  - **Explore practical benefits:** Participant feedback suggested practical benefits and gamification as possible incentives for performing the authentication ceremony. Their effects on verification behavior were not evaluated and should be investigated in future studies.

12.1 AVAILABILITY

Alongside this thesis, we are publishing the full source code of twonly, including all modifications made throughout this work. Furthermore, all study documents (questionnaires, study policies, consent forms, and receipts) and the scripts used to analyze the raw survey and in-app statistical data are available at https://dev.seemoo.tu-darmstadt.de/stud-theses/msc-tobias-mueller.

A

STUDY POLICY (PLURALISTIC WALKTHROUGH)

This study policy is used to standardize the user study sessions and ensure that the moderator stays neutral during the discussions. The policy includes overall guidelines applied during the session and a step-by-step protocol with standard phrases read to the participants.

A.1 GUIDELINES FOR THE WHOLE SESSION

- Stay neutral.
- Do not intervene.
- Protocol anomalies. Note the participant ID on.

A.2 STEP-BY-STEP STUDY PROTOCOL

- Greeting: "Hello! Thank you for agreeing to participate in this study. Today, we will look at different designs together. First, I will show you how to interact with the design tool. Then, you will receive different scenarios and look at various application screens. You will first write down individually on a separate sheet of paper what action you would take. Afterward, I will describe the actual intended action. Then, you will discuss your actions with each other, including why you might have made a different decision and how you would improve the user interface or the text. Following your discussion, I will join the conversation and, among other things, explain why the user interface was structured this way. After that, we will move on to the next scenario or application screens until none are left. The entire session will take about 60 minutes. At the end, everyone will receive a compensation of €15 per person."
- Data collection notice: "To demonstrate the representativeness of the results at the end of the study, some personal information needs to be collected."
- Note the participant IDs on the consent forms and hand them out.
- Request signatures: "This consent form explains exactly how the data will be handled. Please read it carefully and ask if anything is unclear. If you agree, please sign at the end."
- Check the signatures.

- When using the UI prototype:
  - Ask the participants to activate the "Do Not Disturb" mode for the duration of the study so they are not distracted by incoming messages.
  - Show the participants a QR code so they can open the design tool in a browser on their own smartphones.
  - Ensure that the participants have disabled the highlighting of clickable areas.
  - Go through the example together and emphasize that the participants should only proceed when instructed to do so.
  - Ask if everyone has understood how to use the tool.
- When using the presentation:
  - Ensure that everyone can see the presentation and can read the text.
- Request consent for audio recording: "The audio recording mentioned in the consent form facilitates the transcription process. Do you agree to an audio recording starting from this point? The alternative would be for me to write everything down manually."
- If agreed, start the recording.
- State or note the participant IDs for the transcript.
- Note the current time.
- Alleviate potential nervousness: "Today, the designs are being tested, not you. If something seems difficult or does not make sense, that is exactly what needs to be discovered. Please share all your thoughts, whether positive or negative. All comments are helpful for the evaluation."
- Iterate over all scenarios and application screens:
  - If it is a scenario:
    - Ask the participants to read the scenario and confirm if they have understood it.
    - Then proceed to the next application screen.
  - If it is an application screen:
    - Ask the participants to write down the action they would perform.
    - Wait until all participants have noted their independent actions and signaled that they are ready.
    - Read the intended action aloud.

  - Then ask the participants to describe what they did and what their reasoning was.
  - Ask the participants to discuss potential usability problems they encountered or suspect.
  - As the discussion comes to an end, the moderator joins in and explains the reasoning behind the design.
  - The participants and the moderator then discuss how the usability problems could be resolved.
  - After this second discussion concludes, proceed to the next application screen or finish the task rounds.

- When all application screens have been discussed: "Thank you for your participation. There are two more questionnaires for you. Please fill them out. It would be helpful if you could provide the reasons for your decisions on the task questionnaire."
- After the questionnaires are completed, a brief conversation with the moderator follows to answer any remaining questions.
- "Do you have any further thoughts you would like to share?"
- Stop the recording.
- Hand out the compensation and ask for a signature on the receipt.

B

STUDY POLICY (SIMPLIFIED USER TESTING)

To standardize the study sessions and maintain comparability with Kebbedies' studies, the study policy was adapted from Kebbedies to include the onboarding tasks described in Section C. This study policy helps to minimize any influence by the moderator. It includes overall guidelines applied during the session. It also contains a step-by-step protocol with standard phrases read to the participants. The study was performed in the native language of the participants.

B.1 GUIDELINES FOR THE WHOLE SESSION

- Stay neutral.
- Reply to questions with a counter-question.
- Do not intervene.
- Protocol anomalies. Note the participant ID on.

B.2 STEP-BY-STEP STUDY PROTOCOL

- Greeting: "Hello! Thank you for agreeing to participate in our study. Today, you will get an app where you are supposed to perform an activity. We would like to hear your opinion about that. The whole session will take around 60 minutes. In the end, you will receive an expense compensation of €10 per person."
- Notice on data collection: "To show the representativeness of our results at the end, we would like to collect some personal information."
- Write down the participant ID on the consent form and hand it out.
- Ask for signature: "This consent form explains the exact handling of the data. Please read it carefully and ask if something is unclear. If you consent, please put your signature at the end."
- Check signature.
- Ask for consent for audio recording: "The audio recording stated in the consent form makes transcription easier for us. Do you consent to a recording from this point onwards? The alternative is for me to transcribe everything manually."
- If yes, start recording.

- Say or note participant ID for transcript.
- Note the time.
- Think-Aloud: "Now you will get an activity on paper that you are supposed to perform in our app. Imagine that you are performing it in the real world. Please say aloud what you are thinking. If you read something, read it aloud. If you click on something, tell us what you expect to happen."
- Demonstrate thinking aloud using a calculator.
- Minimize possible nervousness: "Today, we are not testing you; we are testing the app. If something seems difficult or does not make sense, that is what we want to find out. Please tell us all your thoughts when you use the app, positive or negative. All remarks are helpful to us."
- Introduce the moderator as a passive participant: "I will take instructions from you and execute them. Please tell me as precisely as possible what I should do. Imagine that I am someone who has never used a smartphone before."
- Read the task aloud and place it in front of the participant.
- If the task was solved or the participant remained on an incorrect path for over ten minutes: "Thank you. I have two more questionnaires for you. Please fill them out. It would be helpful if you could tell us the reasoning behind your choices on the task questionnaire."
- "Do you have any further thoughts you would like to share?"
- Stop recording.
- Hand out compensation and ask for a signature on the receipt.

C

TASKS (SIMPLIFIED USER TESTING)

The initial results and external feedback showed that most participants liked the WoTM approach and had no issues with sharing their mutual friends. Because many non-technical users clicked through the onboarding pages without really reading the text, they were allowed to shorten the onboarding using their own "setup path", which reduced the configuration from seven to only three pages. Users who wanted to customize everything still had that option. To test this, the first task was designed to determine whether users would choose the fast setup or the customized version.

First Task: Onboarding

You have already installed the app and registered.

Your first task is now to complete the onboarding in the app, just as you would normally do. If you are asked for a backup password, use the password: 1234567890. Your first task is officially completed once the onboarding process has been successfully finished and you have reached the chats.

The second task was adapted from Kebbedies to test the improved authentication ceremony with the goal of answering RQ.

Second Task: Perform Authentication Ceremony

We are currently meeting in person and are already connected on the messenger. In the future, we no longer want to meet in person, but communicate digitally and exchange sensitive data.

Your task is to ensure that, after the meeting, we can communicate securely with each other via this app and that no one can read or manipulate our messages. I will remain passive during this activity. If you want me to do something on my smartphone, please give me clear instructions. Let me know clearly when you think you have successfully completed the task.

If a participant chose the fast setup path, they were assigned a third task where they had to complete the customization path. The goal of this task was to understand the participants' perception of the WoTM approach, particularly their opinions regarding its default activation without explicit consent, as the settings were only accessible through the customized setup path.

Third Task: Onboarding (Conditional)

You are in onboarding again, but this time the "Customize" option has been selected.

Your task is now to complete the onboarding in the app again. Your task is officially completed once the onboarding process has been successfully finished and you have reached the chats again.

D

TASK QUESTIONNAIRE (SIMPLIFIED USER TESTING)

After the think-aloud session, this questionnaire, adopted from Kebbedies [16], was given to participants. They were instructed to fill it out and verbally explain their reasoning. The reasoning was recorded in the respective transcript.

Please indicate the degree to which you agree with the following statements (five-point Likert scale):

1. I think that I solved the task successfully.
2. I found it easy to solve the task.
3. I am satisfied with the user experience of the app ().
4. The tested functions of the app meet my requirements for a messenger.
5. I trust that the app is secure.
6. I understand the principle of security of the app.

How likely is it that you would recommend the tested app to friends or colleagues? (0 = Very unlikely to 10 = Very likely)

E

COMPARATIVE ONLINE SURVEY

The following self-administered questionnaire was used to conduct the comparative online survey.

E.1 MESSENGER USAGE COMPARISON

The following questions were presented in a two-column layout, with identical questions asked separately for Signal and twonly.

1. How long have you been using this messenger?
   (*less than half a year, between half a year and one year, 1 year, 2 years, more than 2 years*)
2. Do you know that you can verify other people in this messenger?
   (*yes, no*)
3. Have you ever used this feature?
   (*yes, no*)
4. (*Only if Q3 = yes*) How many of your contacts have you already verified?
   (*none, 1 to 2, 3 to 5, 6 to 10, more than 10*)
5. Why have you verified contacts, or why not?
   free text

E.2 MOTIVATION, IMPORTANCE, AND PREFERENCE

1. Is the verification of contacts a feature that you find relevant for a messenger?
   (*yes, no*) + *free text for reasoning*
2. What would motivate you to perform the verification, and which aspects are particularly important to you?
   free text
3. Which version of the verification would you rather use in everyday life?
   (*Signal, twonly, neither*) + *free text for reasoning*
4. In Signal, the feature is called "safety number comparison"; in twonly, it is called "verification badge." Which term do you find more understandable?
   (*Signal* (*safety number*), *twonly* (*verification badge*), *no difference*) + *free text: why?*
5. In twonly, only one person needs to scan the QR code; the other person is then automatically notified and also verified. How do you rate this compared to Signal,

where both people each need to scan?
free text

6. Which version gave you a greater sense of security in communication, and why?
(*Signal, twonly, no difference*) + *free text*

E.2.1 *Importance of Individual Aspects (Likert Scale 1 to 5)*

"How important are the following aspects to you when performing the verification?" (1 = not important at all, 5 = very important)

1. Duration of the process
2. Simplicity of the procedure
3. Comprehensibility of the workflow
4. Visual feedback (e.g., checkmarks, confirmation)
5. Trust in the security of the process
6. Discoverability of the verification feature

E.3 TWONLY-SPECIFIC: VERIFICATION BADGES

"In twonly, verification badges were introduced in recent months."

1. Did you notice the verification badge?
(*yes, no*)
2. Has the badge changed your use of twonly?
(*yes, no*) + *free text for reasoning*
3. How do you handle the red verification badge?
(*I ignore it, it unsettles me, it motivates me to verify*)
4. Do you have contacts with a blue verification badge? If so, how many? How do you like the feature?
(*none, 1 to 2, 3 to 5, 6 to 10, more than 10*) + *free text*
5. Does the verification status of a contact change how you communicate with that person? (e.g., whether you share more sensitive information)
free text
6. The blue badge in twonly indicates that mutual friends have also verified the contact, without the server being able to create a social graph. Would you want such a feature

in Signal as well?
(*yes, no*) + *free text for reasoning*

7. Would you also like to have the verification badges (red for unverified, blue for verified) in Signal?
(*yes, no*) + *free text for reasoning*

E.3.1 *Agreement with Statements on the Badge Feature (Likert Scale 1 to 5)*

(1 = strongly disagree, 5 = strongly agree)

1. I find the visibility of the verification badges practical.
2. The feature increases my trust in the messenger.
3. The feature rather annoys me.
4. The red badge unsettles me.
5. I understood what the verification badge means.

F

PERSON QUESTIONNAIRE

After participants finished their think-aloud sessions and filled out the task questionnaire, the following person questionnaires were handed out to collect demographics and control variables. This questionnaire was also derived from Kebbedies; the only change is the removal of the participants gender, as this information is not necessary to answer the research question.

F.1 DEMOGRAPHIC DATA

1. How old are you?
   (*18-22, 23-27, 28-32, 33-37, 38-42, 43-47, 48-52, 53-57, 58-62, 63-67, 68-72, 73-77, 77 or older, no answer*)
2. What is your highest educational qualification?
   (*still in school, lower secondary education, polytechnical high school, intermediary secondary education, university entrance qualification, no educational qualification, free text, no answer*)
3. What is your highest vocational qualification?
   vocational training or training in dual system, technical college diploma, technical college diploma in the former GDR, bachelor, master, diploma, PhD, no vocational qualification, free text, no answer
4. Are you currently enrolled in a study program? (Bachelor, Master, Diploma, State examination, Magister)
   (*yes, no, no answer*)
5. If so, which study program?
   free text, no answer
6. Are you currently working?
   (*yes, no, no answer*)
7. If so, what is your field of work?
   (*free text, no answer*)

F.2 TECHNICAL KNOWLEDGE

1. Have you been using a smartphone for more than two years?
   (*yes, no*)

2. How do you estimate your general computer knowledge?
   (*beginner, mid, advanced*)
3. How do you estimate your IT-Security Knowledge?
   (*beginner, mid, advanced*)
4. Which of the following messengers have you used before?
   (*WhatsApp, Signal, Telegram, Facebook Messenger, iMessage, Snapchat, Matrix, others*)

F.3 TECHNICAL AFFINITY

The following deals with your interaction with technical systems. "Technical systems" include Apps, other software applications, and entire digital devices, e.g., smartphones, computers, televisions, and car navigation. (Six-point Likert Scale)

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.

BIBLIOGRAPHY

[1] “Most popular messaging apps 2025.” Accessed: May 11, 2026. [Online]. Available: https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/

[2] “Numerous journalists targeted in attack via Signal Messenger.” Accessed: Jun. 05, 2026. [Online]. Available: https://netzpolitik.org/2026/phishing-attack-numerous-journalists-targeted-in-attack-via-signal-messenger/

[3] heise online, “Nach Signal-Affäre: Auch EU-Minister kommunizieren über Gruppenchat.” Accessed: Jun. 12, 2026. [Online]. Available: https://www.heise.de/news/EU-gesteht-Existenz-von-Signal-Gruppenchat-der-EU-Aussenminister-ein-10353749.html

[4] TheGuardian, “Exclusive: how the Atlantic’s Jeffrey Goldberg got added to the White House Signal group chat | Signal group chat leak | The Guardian.” Accessed: Feb. 22, 2026. [Online]. Available: https://web.archive.org/web/20260131112207/https://www.theguardian.com/us-news/2025/apr/06/signal-group-chat-leak-how-it-happened

[5] “Chat Control: The EU’s CSAM scanner proposal – Patrick Breyer.” Accessed: Aug. 22, 2026. [Online]. Available: https://www.patrick-breyer.de/en/posts/chat-control/

[6] “"Die Leute sind gefickt": Wie die Polizei heimlich ein Handynetzwerk für Drogengangs infiltrierte.” Accessed: Aug. 22, 2026. [Online]. Available: https://www.vice.com/de/article/encrochat-hack-wie-die-polizei-ein-handynetzwerk-fur-drogengangs-infiltrierte/

[7] heise online, “Signal phishing warning: Trigger likely attack on Julia Klöckner.” Accessed: May 06, 2026. [Online]. Available: https://www.heise.de/en/news/Signal-phishing-warning-Trigger-likely-attack-on-Julia-Kloeckner-11268773.html

[8] “Police have achieved a breakthrough in the interception and decryption of crypto communication | politie.nl.” Accessed: Jun. 05, 2026. [Online]. Available: https://web.archive.org/web/20181108124708/https://www.politie.nl/en/news/2018/november/02-apeldoorn-police-have-achieved-a-breakthrough-in-the-interception-and-decryption-of-crypto-communication.html

[9] D. Goodin, “Police decrypt 258,000 messages after breaking pricey IronChat crypto app.” Accessed: Jun. 05, 2026. [Online]. Available: https://arstechnica.com/information-technology/2018/11/police-decrypt-258000-messages-after-breaking-pricey-ironchat-crypto-app/

[10] “What is a safety number and why do I see that it changed?.” Accessed: Sep. 08, 2026. [Online]. Available: https://support.signal.org/hc/en-us/articles/360007060632-What-is-a-safety-number-and-why-do-I-see-that-it-changed

[11] “Deploying key transparency at WhatsApp.” Accessed: Jun. 05, 2026. [Online]. Available: https://engineering.fb.com/2023/04/13/security/whatsapp-key-transparency/

[12] "Automatic Key Verification." Accessed: Jun. 05, 2026. [Online]. Available: https://support.signal.org/hc/en-us/articles/10223569377562-Automatic-Key-Verification

[13] K. Fischer, M. Keil, A. Buckmann, and M. A. Sasse, "'If You Want to Encrypt It Really, Really Hardcore...': User Perceptions of Key Transparency in WhatsApp," *Proceedings on Privacy Enhancing Technologies*, 2025, Accessed: Jun. 05, 2026. [Online]. Available: https://petsymposium.org/popets/2025/popets-2025-0170.php

[14] E. Vaziripour *et al.*, "Is that you, Alice? A Usability Study of the Authentication Ceremony of Secure Messaging Applications," 2017, pp. 29–47. Accessed: Dec. 06, 2025. [Online]. Available: https://www.usenix.org/conference/soups2017/technical-sessions/presentation/vaziripour

[15] M. Fassl, "Usable authentication ceremonies in secure instant messaging," Thesis, 2018. doi: 10.34726/hss.2018.47902.

[16] M. Kebbedies, "Usable Authentication for End-to-End Encrypted Messengers," 2024.

[17] M. Fassl and K. Krombholz, "Why I Can't Authenticate — Understanding the Low Adoption of Authentication Ceremonies with Autoethnography," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, in CHI '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–15. doi: 10.1145/3544548.3581508.

[18] I. Nielsen, "Usability engineering at a discount," in *Proceedings of the Third International Conference on Human-Computer Interaction on Designing and Using HumanComputer Interfaces and Knowledge Based Systems (2nd Ed.)*, 1989, pp. 394–401.

[19] "Specifications >> The Sesame Algorithm: Session Management for Asynchronous Message Encryption." Accessed: Feb. 17, 2026. [Online]. Available: https://signal.org/docs/specifications/sesame/

[20] (@signalapp@mastodon.world) Signal, "However, sophisticated attackers have engaged in a harmful phishing campaign, posing as "Signal Support" by changing their profile display...." Accessed: May 06, 2026. [Online]. Available: https://mastodon.world/@signalapp/116478660038777409

[21] "Discount Usability: 20 Years." Accessed: Apr. 08, 2026. [Online]. Available: https://www.nngroup.com/articles/discount-usability-20-years/

[22] "Pluralistic Usability Walkthrough." Accessed: Mar. 15, 2026. [Online]. Available: https://www.usabilitybok.org/pluralistic-walkthrough

[23] R. Bias, "Interface-Walkthroughs: efficient collaborative testing," *IEEE Software*, vol. 8, no. 5, pp. 94–95, Sep. 1991, doi: 10.1109/52.84220.

[24] "Usability Inspection Methods: Book by Jakob Nielsen." Accessed: Mar. 15, 2026. [Online]. Available: https://www.nngroup.com/books/usability-inspection-methods/

[25] "Iterative Design of User Interfaces." Accessed: Apr. 08, 2026. [Online]. Available: https://www.nngroup.com/articles/iterative-design/

[26] E. Vaziripour *et al.*, "Action Needed! Helping Users Find and Complete the Authentication Ceremony in Signal," 2018, pp. 47–62. Accessed: Dec. 06, 2025. [Online]. Available: https://www.usenix.org/conference/soups2018/presentation/vaziripour

[27] M. Fassl, L. T. Gröber, and K. Krombholz, "Exploring User-Centered Security Design for Usable Authentication Ceremonies," in *Proceedings of the 2021 CHI Confer-*

ence on Human Factors in Computing Systems, in CHI '21. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–15. doi: 10.1145/3411764.3445164.

[28] M. Häring, J. A. Grohs, E. Tiefenau, M. Smith, and C. Tiefenau, "Can Johnny be a whistleblower? A qualitative user study of a social authentication Signal extension in an adversarial scenario," 2024, pp. 259–278. Accessed: Jan. 18, 2026. [Online]. Available: https://www.usenix.org/conference/soups2024/presentation/haring

[29] "penpot/penpot." Accessed: Apr. 08, 2026. [Online]. Available: https://github.com/penpot/penpot

[30] "twonlyapp/twonly-app." Accessed: May 06, 2026. [Online]. Available: https://github.com/twonlyapp/twonly-app

[31] "signalapp/libsignal." Accessed: Aug. 19, 2026. [Online]. Available: https://github.com/signalapp/libsignal

[32] "Why You Only Need to Test with 5 Users." Accessed: Apr. 08, 2026. [Online]. Available: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

[33] D. Wessel, C. Attig, and T. Franke, "ATI-S - An Ultra-Short Scale for Assessing Affinity for Technology Interaction in User Studies," in *Proceedings of Mensch und Computer 2019*, in MuC '19. New York, NY, USA: Association for Computing Machinery, Sep. 2019, pp. 147–154. doi: 10.1145/3340764.3340766.

[34] A. J. Nederhof, "Methods of coping with social desirability bias: A review," *European Journal of Social Psychology*, vol. 15, no. 3, pp. 263–280, 1985, doi: 10.1002/ejsp.2420150303.

[35] J. G. Adair, "The Hawthorne effect: A reconsideration of the methodological artifact," *Journal of Applied Psychology*, vol. 69, no. 2, pp. 334–345, 1984, doi: 10.1037/0021-9010.69.2.334.

[36] "twonly.eu (@twonly@mastodon.social)." Accessed: Aug. 19, 2026. [Online]. Available: https://mastodon.social/@twonly/115966775370449880

[37] (@tsmreu@infosec.exchange) Tobias, "Du verwendest Signal und twonly seit mehr als einem Monat?Dann würde ich mich sehr freuen, wenn du 5 Minuten Zeit hättest und an meiner...." Accessed: Aug. 19, 2026. [Online]. Available: https://infosec.exchange/@tsmreu/117036602826559240

[38] kaixxx, "kaixxx/noScribe." Accessed: Apr. 09, 2026. [Online]. Available: https://github.com/kaixxx/noScribe

[39] F. Putz, S. Haesler, and M. Hollick, "Sounds Good? Fast and Secure Contact Exchange in Groups," *Proc. ACM Hum.-Comput. Interact.*, vol. 8, no. CSCW2, p. 425:1–425:44, Nov. 2024, doi: 10.1145/3686964.

[40] M. Reuter, "Phishing attack: Numerous journalists targeted in attack via Signal Messenger." Accessed: Mar. 31, 2026. [Online]. Available: https://netzpolitik.org/2026/phishing-attack-numerous-journalists-targeted-in-attack-via-signal-messenger/

[41] "Deeplinks | App architecture." Accessed: Mar. 31, 2026. [Online]. Available: https://developer.android.com/training/app-links?hl=de

[42] "Allowing apps and websites to link to your content." Accessed: Mar. 31, 2026. [Online]. Available: https://developer.apple.com/documentation/xcode/allowing-apps-and-websites-to-link-to-your-content/

[43] T. Müller, "twonly – Finding Friends Without Phone Numbers - The privacy-friendly, secure, European alternative." Accessed: Jul. 16, 2026. [Online]. Available: https://twonly.eu/en/blog/2026-mutual-friends
[44] Y. Gokce, "The Bylock fallacy: An In-depth Analysis of the Bylock Investigations in Turkey," *Digital Investigation*, vol. 26, pp. 81–91, Sep. 2018, doi: 10.1016/j.diin.2018.06.002.
[45] "Signal (@signalapp@mastodon.world)." Accessed: Jul. 09, 2026. [Online]. Available: https://mastodon.world/@signalapp/116558180769475668
[46] "Keep your phone number private with Signal usernames." Accessed: Sep. 21, 2026. [Online]. Available: https://signal.org/blog/phone-number-privacy-usernames/
[47] J. Dykstra, "Invisible Security: Protecting Users with No Time to Spare," in *2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC)*, Dec. 2020, pp. 184–190. doi: 10.1109/CIC50333.2020.00031.
[48] C. Reuter, L. L. Iacono, and A. Benlian, "A quarter century of usable security and privacy research: transparency, tailorability, and the road ahead," *Behaviour & Information Technology*, vol. 41, no. 10, pp. 2035–2048, Jul. 2022, doi: 10.1080/0144929X.2022.2080908.
[49] "Technology preview: Private contact discovery for Signal." Accessed: Sep. 21, 2026. [Online]. Available: https://signal.org/blog/private-contact-discovery/
[50] T. Müller, "twonly – Introducing Passwordless Backups - The privacy-friendly, secure, European alternative." Accessed: Jul. 16, 2026. [Online]. Available: https://twonly.eu/en/blog/2026-passwordless-backup

AI DECLARATION

This thesis was written independently and was linguistically revised with the help of Claude Opus 5 and GPT-5.6 Sol.

ERKLÄRUNG ZUR ABSCHLUSSARBEIT

gemäß § 22 Abs. 7 und § 23 Abs. 7 APB TU Darmstadt

Hiermit versichere ich, Tobias Müller, die vorliegende Master's Thesis gemäß § 22 Abs. 7 APB der TU Darmstadt ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Mir ist bekannt, dass im Falle eines Plagiats (§ 38 Abs. 2 APB) ein Täuschungsversuch vorliegt, der dazu führt, dass die Arbeit mit 5,0 bewertet und damit ein Prüfungsversuch verbraucht wird. Abschlussarbeiten dürfen nur einmal wiederholt werden. Bei der abgegebenen Thesis stimmen die schriftliche und die zur Archivierung eingereichte elektronische Fassung gemäß § 23 Abs. 7 APB überein.

THESIS STATEMENT

pursuant to § 22 paragraph 7 and § 23 paragraph 7 of APB TU Darmstadt

I herewith formally declare that I, Tobias Müller, have written the submitted Master's Thesis independently pursuant to § 22 paragraph 7 of APB TU Darmstadt. I did not use any outside support except for the quoted literature and other sources mentioned in the paper. I clearly marked and separately listed all of the literature and all of the other sources which I employed when producing this academic work, either literally or in content. This thesis has not been handed in or published before in the same or similar form. I am aware, that in case of an attempt at deception based on plagiarism (§ 38 paragraph 2 APB), the thesis would be graded with 5.0 and counted as one failed examination attempt. The thesis may only be repeated once. In the submitted thesis the written copies and the electronic version for archiving are pursuant to § 23 paragraph 7 of APB identical in content.

Darmstadt, September 30, 2026

Tobias Müller