Evaluating In-Workflow Messages for Improving Mental Models of End-to-End Encryption

Omer Akgul, Wei Bai, Shruti Das, and Michelle L. Mazurek, *University of Maryland*

https://www.usenix.org/conference/usenixsecurity21/presentation/akgul

This paper is included in the Proceedings of the
30th USENIX Security Symposium.

August 11–13, 2021

978-1-939133-24-3

Open access to the Proceedings of the
30th USENIX Security Symposium
is sponsored by USENIX.

Evaluating In-Workflow Messages for Improving Mental Models of End-to-End Encryption

Omer Akgul, Wei Bai, Shruti Das, and Michelle L. Mazurek
University of Maryland

Abstract

As large messaging providers increasingly adopt end-to-end encryption, private communication is readily available to more users than ever before. However, misunderstandings of end-to-end encryption's benefits and shortcomings limit people's ability to make informed choices about how and when to use these services. This paper explores the potential of using short educational messages, built into messaging workflows, to improve users' functional mental models of secure communication. A preliminary survey study (n=461) finds that such messages, when used in isolation, can effectively improve understanding of several key concepts. We then conduct a longitudinal study (n=61) to test these messages in a more realistic environment: embedded into a secure messaging app. In this second study, we do not find statistically significant evidence of improvement in mental models; however, qualitative evidence from participant interviews suggests that if made more salient, such messages could have potential to improve users' understanding.

1 Introduction

Recent adoption of end-to-end encryption (e2e encryption) by popular messaging apps such as iMessage, Facebook Messenger, and Whatsapp [6,22,23] has enabled strong communications privacy protections for billions of users globally [69].

People, however, often fail to use e2e encrypted apps confidently and correctly [2,27]. Often this is because they misunderstand key aspects of e2e encryption, sometimes believing it protects from broad classes of "hacking" [2], and sometimes believing any attempt at ensuring privacy is hopeless in the face of powerful or skilled adversaries [2,3,17,27]. Users also struggle to choose appropriate communication mechanisms for sharing private information [2,3,27].

For people to make good decisions about their communications — including opting for more private communications when appropriate — they need to develop strong *functional mental models* of secure communications. Mental models refer to a user's understanding of how a system works, its inputs and outputs, and other effects that can be expected from using it. Functional models, specifically, are not directly about how a system works but rather help a user to understand when to use a system and predict how it will behave [19]; as such, they need not be fully correct in all details, but must be sufficient to be useful. Unfortunately, existing functional models of secure communication are often inadequate [2,17,36,40].

Various attempts at influencing these mental models exist both in the research literature and in the broader privacy community. Organizations like the Electronic Freedom Foundation, Citizen Lab, and the Library Freedom Project have produced broad guidance for improving personal privacy and security, including discussion of secure messaging [20,21,42]. Researchers have tested a variety of metaphors for better explaining encryption, with only limited success [18,60]. Others have worked to make authentication ceremonies more usable, again with mixed results [56,63,64,72].

In prior work, we were able to meaningfully improve users' understanding by asking them to complete a brief tutorial as part of a larger experiment on secure messaging preferences [8]. However, it is of course not realistic to expect many users to sit down and participate in a tutorial when they are not being paid as experimental participants.

This raises the question of whether it is possible to convey some of the key information that might be included in such a tutorial more naturally, for example during splash or interstitial screens within a messaging app, or as reminders inserted automatically by the software during a text conversation. This approach builds on a long tradition in the human-centered security community of nudging users toward privacy-protective behaviors [4,5,12,37,41,65,66,72].

In this paper, we investigate the potential of in-workflow educational messages to help users improve their functional mental models of e2e encryption and therefore their communications decision-making. To this end, we design a series of messages with different lengths, emphasizing different key principles related to e2e encryption. These messages build directly on our prior work exploring which e2e encryption concepts are most important and useful to convey to users [9].

We preliminarily test these messages using an online survey study with lay users (n=461) and find the messages are generally effective at improving participants' understanding. Longer messages are most effective, while shorter messages can successfully convey specific points, at the possible risk of enabling misunderstandings of concepts that aren't covered.

We then embed our short messages into a chat app (adapted from Signal) and ask lay users (n=61) to use the app daily for about three weeks. Unfortunately, we find no statistical evidence that these embedded messages effectively improved mental models. A post-study interview (n=19) suggests that although many participants noticed our messages, they did not pay attention to them, limiting their impact.

Overall our results, while somewhat disappointing, suggest short educational messages can be useful if users pay attention to them. This provides some hope that if in-workflow messages are made more salient (and perhaps somewhat more intrusive), they may have potential to improve mental models and support better communications privacy decision-making.

2 Related work

We discuss related work in three key areas.

Usability, adoption, and mental models of encrypted communication tools For more than 20 years, researchers have been exploring the usability and adoption of encrypted communications tools. Extensive studies of encrypted email tools have identified a range of issues that inhibit adoption and use, including challenges in key management, complex interfaces, social and cultural factors, network effects, and user misunderstandings [8, 24–26, 47–50, 53].

The incorporation of e2e encryption into centralized secure-messaging apps has reduced the salience of key management, although several researchers have documented remaining challenges in authenticating keys [56, 64]. Researchers have demonstrated that network effects play a large role in inhibiting adoption of security-focused messaging apps [3, 36], but the integration of e2e encryption into already-popular apps such as WhatsApp and iMessage has to a large extent overcome this problem [69].

Adoption, however, has not proven entirely sufficient. In interviews and surveys, researchers find that many users do not believe that e2e encrypted tools provide meaningful protection [2, 3, 17, 27]. As a result, people frequently make less than optimal choices about how to communicate sensitive information — such as preferring SMS messaging [3, 17] — or use ad-hoc protection strategies [27].

These misconceptions appear to arise from incorrect or imprecise mental models of encrypted communication, such as beliefs that anyone with computer-science knowledge or who knows an encryption algorithm can decrypt its results [3, 17], misunderstandings of the role of service providers in communication paths [17], and distinctions between e2e encryption and other kinds of communications [3, 17, 40]. These misunderstandings reflect broader inaccuracies in mental models of encryption generally [73], as well as common beliefs regarding security or privacy generally that "ordinary" people are not important or valuable enough to be targeted [67]. In this paper, we measure changes in mental models in part by looking specifically for misunderstandings and key concepts identified in the prior works cited here.

Nudging security and privacy behaviors Attempts to integrate security or privacy warnings or messages into UI elements and workflows, prompting more secure or private behavior, are sometimes known as *nudging* [4, 41]. Examples include improving feedback during password creation (e.g., [61]), during software updates (e.g., [38]), and during semi-automated checks for malware [12]. Nudging has also been used to promote privacy-preserving behavior in the context of social media [37, 65, 66] and mobile app permissions [5], and even correct use of authentication ceremonies in encrypted messaging [72].

Rather than prompting specific behaviors, our in-workflow messages are intended to improve users' understanding and functional mental models. In a sense, we seek to improve on resources such as existing tooltips in e2e encrypted messaging apps, which have been shown to be ineffective [17].

Teaching encryption and secure communication Researchers have experimented with a variety of encryption metaphors, none of which has as yet been highly successful [18, 60]. Although we do use a lock-key metaphor in some of our messages, we mainly follow results from our prior work which suggest focusing primarily on functional models — what e2e encryption can and cannot do — rather than on structural information about how e2e encryption works [9].

Researchers have also focused attention on improving users' understanding of and facility with authentication ceremonies [56, 62, 63, 72]. Because it has been covered in depth previously, we do not address authentication ceremonies in this work. However, our work does use similar methods and could be combined with prior work on ceremoneis [72] to provide a more complete view of e2e encryption.

Nonprofit and advocacy organizations have produced blog posts, interactive guides, infographics, and other educational materials related to personal privacy and security, including secure communications [20, 21, 42]. These tutorials serve a different niche than our work; they focus on aiding (in detail) people who seek out guidance in privacy and secure messaging, while we target small, high-level improvements for casual users. We do, however, incorporate some concepts from these tutorials into the messages we design.

We build directly on our prior work using tutorials and participatory design to explore which concepts and explanations related to e2e encryption are most important, surprising, and challenging for users [9]. Those results suggest focusing on confidentiality and explicit discussion of risks, while noting

that certain misconceptions are difficult to overcome. Educational messages in this study were based on these findings.

3 Survey Study: Methods

We first conducted an online, between-subjects study to preliminarily measure the effectiveness of brief educational messages for e2e encryption novices. As with our prior work, our messages emphasize actionable information about threats, non-threats, and appropriate usage of e2e encryption, rather than focusing on how encryption works [9].

We recruited participants from the Prolific.[1] After consent, participants were directed to an online survey consisting of five parts. First, we asked background questions about self-reported knowledge of technical and encryption concepts as well as general web-use concepts [29], in order to filter out participants with too much expertise. Second, we introduced *TextLight*, a hypothetical e2e encrypted application.[2] Third, to obtain a baseline of participants' mental models, they completed a pre-intervention questionnaire (described in Section 3.1). Fourth, they viewed one educational message we created (detailed in Section 3.2). Fifth, they completed a post-intervention questionnaire containing the same questions as the pre-intervention questionnaire, allowing us to measure differences. Finally, the participants answered demographic questions. Throughout the study, we referenced e2e encryption in the context of TextLight, to make the concept concrete[3].

Our study protocol was approved by the University of Maryland Institutional Review Board (IRB).

3.1 Communications privacy questionnaire

We investigate knowledge of privacy threats by asking questions about the capabilities of various adversaries (see Table 1). Participants were asked, on a 5-point Likert scale (strongly disagree to strongly agree), "Based on your understanding of TextLight's end-to-end encryption, please indicate whether you agree or disagree that [ADVERSARY] has/have the following abilities, regardless of their motivation to do so." We use Likert scales to enable detection of smaller shifts in mental models (compared to binary-choice questions). Similar methodology has been shown to be effective in prior work [9,46]. We asked this question for every combination of the adversaries and capabilities listed in Table 1.

The adversaries and capabilities we selected were mainly adopted from prior work [2,9,18] and in many cases reflect real-world examples of privacy breaches [1, 10, 16, 30, 70]. The adversaries roughly fit into three top level categories: (1) endpoint adversaries, (2) communication providers, and (3) outsiders who might be capable of intercepting communications. The capabilities capture privacy attacks e2e encryption can and cannot protect against. They address confidentiality, integrity, and authenticity, as well as the capability to learn metadata. Finally, to address differences between e2e encrypted and non-e2e encrypted tools, we ask participants to rate the chance of each adversary reading the contents of a message if TextLight *were not* e2e encrypted.

After the adversary/capability questions each participant was asked one free-response question about why they gave their specific answer for one adversary/capability pair, chosen at random per participant. We use these responses as an attention check and to validate that our participants were interpreting the questions as we had intended.

Adversary	Description (as it appeared in the questionnaire)
Employee	People employed by TextLight.
ISP	Your mobile service provider (Verizon, AT&T, Sprint, etc.).
Hacker	Hackers who have compromised the TextLight servers.
Government	A government intelligence or national security agency.
Unlocked Phone	Someone who has access to your unlocked phone.
Malware	Someone who has successfully installed malware on your phone.

Capability	Description (as it appeared in the questionnaire)
Read	Can see what is in the message you have sent on TextLight.
Change	Can change what is in the message after it is sent through the TextLight app. This means the person you are texting with may receive a different message than the one you have sent.
Impersonate	Can pretend to be you on the TextLight app to send messages to other people in your name.
Metadata	Can see that you have sent a message on TextLight, without knowing the content of the message.
Not-E2EE	If TextLight **IS NOT** end-to-end encrypted, can see what is in the message you have sent on TextLight.

Table 1: Adversaries and capabilities used in the communications privacy questionnaire. Participants were asked about all adversary-capability pairs. We refer to the combination of read, change, and impersonate as the *interception* capability.

3.2 Educational messages

When creating new educational messages about end-to-end encryption, we first surveyed existing messages from academia [9,18,72,73] and industry [6,23,43,44,52,54,57,59,71]. We extracted key concepts from these materials and synthesized them into five principles:

[1] https://www.prolific.co

[2] As with prior work ([2,9]) we use a hypothetical application name to avoid confounds related to participants' trust in different brands [33].

[3] Survey text can be found in the extended paper; see Appendix D

- **Confidentiality**. The most frequently mentioned concept in the prior work we reviewed; we previously found it the most important aspect to convey [9]. We aimed to explain that e2e encryption protects the content of messages from adversaries between the sender and intended recipient.
- **Risks**. Risk communication has been shown to be effective both in computer security broadly [7, 13] and secure messaging specifically [9, 72]. Our previous work suggests specific risks were both important and surprising to users [9]. We aim to point out specific adversaries and their capabilities, with an emphasis on comparing risks with and without e2e encryption.
- **Mechanism**. Our goal here was not to communicate technical details, but rather convey a simplified structural model of e2e encryption to support our key functional concepts [19]. Our previous work suggested this kind of information can be useful to certain users when kept brief and focused on confidentiality [9]. As there is no consensus about the "best" metaphor, we adopted the key-lock metaphor [9, 60].
- **Endpoint weakness**. Several real-world privacy breaches have demonstrated the endpoint weakness of e2e encrypted systems [1, 16, 70]. Our prior research indicates that users found weaknesses of e2e encryption to be important [9]. We aim to convey that e2e encryption can't protect against adversaries who have endpoint access, e.g., by installing malware or possessing an unlocked phone.
- **Metadata weakness**. Metadata weaknesses rarely receive attention in e2e encrypted application descriptions and were not emphasized in our prior work; however, we think conveying metadata risks is an important piece of a strong functional model of e2e encryption. We aim to convey that adversaries who cannot access message content may still have access to metadata or infer that a user is communicating using TextLight.

As our goal is to develop messages for integration into existing app workflows, we consider three message lengths that could fit into workflows in different ways:

- **Short**. Designed to fit as an extra message within a chat window (similar to the WhatsApp notification that a chat is e2e encrypted), or fit on a splash screen or interstitial within an app. We designed five Short messages, one for each of the principles described above. (Hereafter, we reference them as s.[principle], named for the principles they embody; for example, s.endpt is a short message referencing endpoint weaknesses.)
- **Medium**. Designed to fit in a popup message if a user clicks on a short message to learn more, or to be included in a summary displayed in an app store. We designed two Medium messages, each of which includes four of the five principles. We left out principles that appeared least effective during pilots (see Section 3.4). We refer to these as m1 (leaves out confidentiality) and m2 (leaves out endpoints).
- **Long**. Designed to be shown on an app's website, or when a user wants to seek out more detailed information. We designed one Long message that includes all five principles; key phrases are highlighted.

In order to accurately measure changes in mental models, we also tested one Control message. This message, adapted from a Telegram description, describes TextLight but does not mention any privacy or security features [58].

Each participant viewed exactly one of these nine messages. The message text is given in Appendix A. More detail about how these messages were derived from our prior work is given in Appendix C.

3.3 Data analysis

Our main analysis goal is to measure the effectiveness of the designed educational messages (especially with respect to control message) in changing mental models of e2e encryption. As such, our main unit of analysis is the *difference* in response to each statement in the communications privacy questionnaire (Table 1), calculated by converting the Likert responses to numeric values 1-5 and subtracting each pre-intervention response from the associated post-intervention response.

Grouping questions To increase reliability of our mental model measurement [28], and to reduce redundant statistical testing on potentially highly correlated questions in the communications privacy questionnaire, we attempt to combine questions about related adversaries and/or capabilities (groupings shown in Table 1). We consider score differences from participants who saw the Long message and use Cronbach's α to test whether questions are correlated. If grouping succeeds ($\alpha > .8$, considered "good" [28]), we average differences across questions in the group to create a single overall difference score.

Attempting to create three adversary groups based on our predefined categories did not yield good internal consistency. However, combining the read, change, and impersonate capabilities for each adversary did achieve good consistency,[4] so we group these questions as the ensemble *interception* capability.[5] This results in six adversaries with three capabilities each (a 40% reduction in statistical testing): interception, metadata, and not-e2e encryption.

Comparing educational messages We employ the following strategy to further reduce unnecessary statistical testing: For each question (or group of combined questions), we

[4] α's are 0.82, 0.84, 0.92, 0.90, 0.92, and 0.90 for Employee, ISP, Server hacker, Government, Unlocked phone, and Malware adversaries respectively.

[5] For the Unlocked-Phone adversary, we group only read and impersonate, as changing messages does not make much sense for this adversary.

(1) calculate the Kruskal Wallis omnibus (KW) test with difference scores as dependent variable and the nine message versions as independent variables. If the KW is significant, we (2) use a two-tailed pairwise Mann-Whitney-U test (MWU) to compare difference scores between Long (our best attempt at explaining e2e encryption) and Control. This comparison indicates whether our messaging is better than no messaging.

If this comparison is significant, we investigate the remaining message versions by (3) computing pairwise MWU's between Long and all other versions, as well as between Control and all other versions. We adjust the resulting p-values for multiple comparison with Holm-Bonferroni correction [32].

We report effect sizes using location-shift estimates [31], which roughly approximate the difference (in Likert-scale points) between the pre- and post-intervention scores. We use a significance level of $\alpha \leq 0.05$ for all statistical tests.

3.4 Pilot studies

We ran two pilot studies prior to the deployment of our survey. We used an initial (partially in-person, pre-COVID-19) pilot with 16 people, recruited through friends and acquaintances, to refine the survey structure and questions.

We used a second pilot on Prolific (n=32), with Short messages only, to refine and validate the survey questions and flow. In addition, results from this pilot informed the choice of principles to include in the Medium messages.

3.5 Limitations

Our controlled experiment provides high internal validity. It approximates a best-case scenario, in which participants are directly instructed to pay attention to the educational message and then asked about it immediately afterward. This allows us to compare messages to each other; however, it does not effectively capture how we expect people to encounter messages in the real world. We use the app study (Section 5) to test the messages with greater ecological validity.

As with similar online studies, our experiment is likely affected by sampling and demand effects. For convenience, and to reduce variability, we limit our sample to the U.S. Typically for Prolific, our sample is not entirely representative of the U.S. population. These limitations reduce generalizability to broader classes of messaging app users.

Demand effects — in which participants report what they think the researchers want to hear — could affect participants' answers, but in this case responding "correctly" indicates the participant has likely learned something. Further, the communications privacy questionnaire might be affecting mental models by prompting users to think critically about e2e encryption. We mitigate this by comparing our experimental groups to a control message. All of these limitations are consistent across conditions, enabling comparison.

Finally, non-parametric statistical tests such as those we use are most appropriate for Likert-type questions, but they have less power than their parametric counterparts, meaning that we may fail to find evidence for small effects.

4 Survey Study: Results

We next detail the results of the survey study.

4.1 Participants

In September 2019, we used Prolific to recruit 578 U.S. residents who do not have programming skills (a proxy for tech-savviness). We discarded 76 participants (13.1%) who reported being comfortable with explaining ("agree" or "strongly agree") end-to-end or symmetric-key encryption. To ensure data quality, we also discarded responses from 12 participants who gave unrelated or nonsensical answers to free-response questions (2.1%). For ease of analysis, we discarded responses from any participant who did not answer all communications privacy questionnaire questions in both the pre- and post-intervention questionnaires (n=29, 5.0%). We analyze responses from the remaining 461 participants. Participants were randomly distributed among message conditions, with twice as many participants allocated to the Long condition because we used it as a basis for our preliminary analysis. After filtering, the Long condition had 92 participants; the other eight conditions had between 42 and 52 each.

On average, participants took just under 10 minutes to complete the study and were compensated $2.00, for an average wage of $12.16/hour.

Table 2 details our participants' demographics. As expected, the sample is younger, whiter, more Asian, and more educated than the U.S. population,[6] but it does capture a broad range of demographics.

4.2 Comparing message versions

We find that the educational messages work significantly better than Control with many adversary-capability pairs. Specifically, Long works best against Control overall, Medium is similar, and Short messages are particularly effective in conveying specific points. However, many participants already had accurate mental models for some aspects of e2e encryption, resulting in no improvement, and we find evidence some of the Short messages may oversell e2e encryption.

We expect participants to learn that Employee, ISP, Government, and Server Hacker adversaries are *less* capable of interception attacks (negative difference scores), while the endpoint adversaries Unlocked Phone and Malware are *more*

[6] https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2018/

		Survey n=461	App n=61
Gender	Female	63.8%	60.7%
	Male	34.9%	36.1%
	Other	1.3%	3.3%
Age	18-24	22.8%	23.0%
	25-29	20.0%	23.0%
	30-39	31.7%	34.4%
	40-49	11.1%	6.6%
	50+	14.3%	13.1%
Ethnicity	White	72.9%	73.0%
	Black or African Am.	8.5%	9.5%
	Asian or Asian Am.	8.5%	11.1%
	Hispanic or Latino	7.2%	16.4%
	Other or mixed race	2.4%	4.0%
Education	Completed H.S. or below	14.1%	9.8%
	Some college, no degree	24.1%	14.8%
	Associate's degree	10.8%	6.6%
	Bachelor's degree	33.4%	37.7%
	Master's degree or higher	14.3%	29.5%
IT-related Job or Degree	Yes	3.3%	9.8%
	No	94.4%	90.2%
	Prefer not to answer	2.4%	0.0%

Table 2: Participant demographics for both studies. Percentages may not add to 100% due to "other" categories and item non-response.

	↗	Long	Medium		Short				
			m1	m2	s.conf	s.meta	s.endpt	s.mech	s.risk
Emp.	Long	—				1.00	1.00		1.00
	Control	-1.00	-1.00	-0.67	-0.33			-0.33*	
ISP	Long	—				0.33	0.67		0.33*
	Control	-0.67		-0.33	-0.33			-0.33	
Gov.	Long	—				0.67	0.67		
	Control	-1.00	-1.00	-0.67	-1.00			-0.67	-0.67
Malware	Long	—		-1.00	-1.00	-0.67		-1.00	-1.00
	Control	0.67	0.67				1.00		

(a) Interception capability location-shift estimates.

	↗	Long	Medium		Short				
			m1	m2	s.conf	s.meta	s.endpt	s.mech	s.risk
ISP	Long	—			-1.00		-1.00	-1.00	-1.00
	Control	0.00*	1.00			1.00			

(b) Metadata capability location-shift estimates.

Better than Control Worse than Long

Table 3: Location-shift estimates from the survey study, measuring change from the message in the row to the message in the column. Populated cells are significant (MWU). * indicates $p \leq 0.002$; $p \leq 0.001$ otherwise. Darker colors denote stronger effects, where red/orange means the message performs worse than Long, and blue/green means it performs better than Control. No messages performed significantly better than Long or worse than Control.

capable of interception attacks (positive). We expect all adversaries to be perceived as *more* capable (positive) of metadata and not-e2e encryption attacks.

Significant results from our condition comparisons are shown in Table 3. We show the distribution of the differences for each adversary-capability pair using violin plots. For additional context, we also plot pre- and post-intervention Likert responses for each. Plots for selected pairs are shown in Figure 1; all plots are available in the extended paper (see Appendix D).

4.2.1 Long is often better than control

Long performs better than Control for several adversary-capability pairs (MWU, $p \leq 0.05$), including the Employee, ISP, Government, and Malware interception capabilities, as well as the ISP metadata capability. The location-shift estimates — that is, how much more effective Long was than Control, expressed in Likert points — range from 0.67 (interception capability of Malware, *more* capable) to -1 (interception capability of Employee and Government, *less* capable).

4.2.2 Some models are already correct

The remaining adversary-capability pairs — Unlocked Phone and Hacker interception, all non-ISP metadata, and all not-e2e encryption — show no significant difference between Long and Control. Many, including all Unlocked Phone and not-e2e encryption, show ceiling effects: participants already had accurate mental models for these questions, leaving little room to observe improvement. One example (Government, not-e2e encryption) is shown in Figure 1, bottom.

4.2.3 Short messages can convey a specific point

Short messages generally work better than control, particularly (as expected) for adversary-capability pairs they directly target. As a reminder, we compare Short messages to Long and Control messages only if the omnibus test comparing all message versions is significant and Long (our best explanation attempt) significantly differs from control.

Better than control but not Long Short versions that aim to give a brief overview of e2e encryption (s.conf and s.mech) perform significantly better than control but not as well as long for the interception capability (see effect sizes in Table 3a). We see this effect for the Employee, ISP, and Government adversaries.

Better than other messages for specific targets We also find that short messages targeting a specific adversary or capability tend to perform well on those questions. For instance, the Short message that targets metadata weakness (s.meta) offers more improvement compared to Control (in terms of

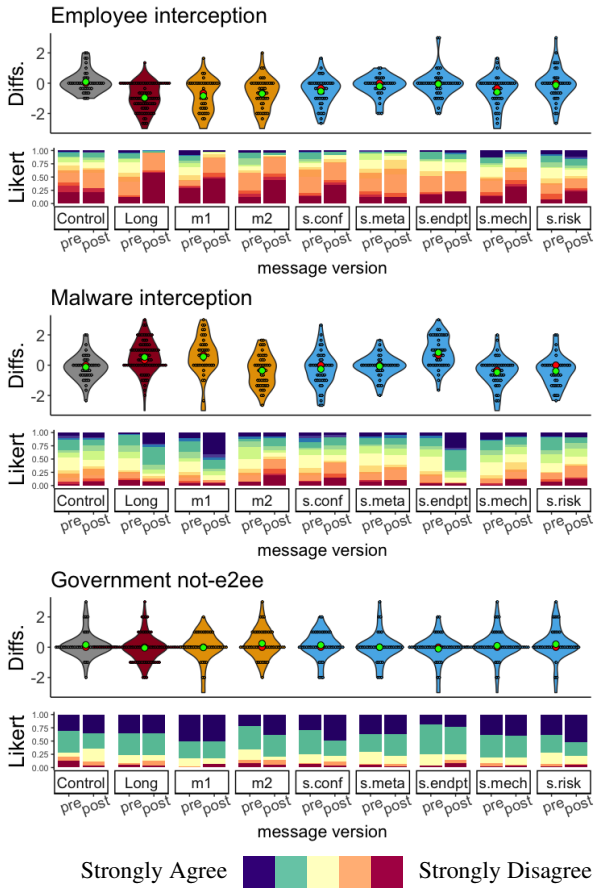


Figure 1: Differences (violin plots: red dots are median, green dots are mean) and pre/post communications privacy questionnaire Likert responses (bar charts) for selected adversary-capability pairs. If the intervention is effective, we expect Employee interception to shift negatively and the others to shift positively. For Employee and Malware interception, several messages improve over Control; for Government not-e2ee encryption, we see a ceiling effect with little room for improvement.

effect size) than Long does. Similarly, the short message that warns against endpoint adversaries (s.endpt) is more effective (in terms of effect size compared to Control) than any other message for the Malware interception capability.

4.2.4 Medium: Better than Control, similar to Long

As with Short, Mediums were only compared to Long and Control if the corresponding Long vs. Control comparisons were significant. Mediums are generally similar to Long both in which results are significant and in effect size. For the interception capability of Employee, ISP, Government, and Malware adversaries, as well as the ISP-metadata pair, at least one Medium is significantly better than Control, with similar effect sizes to Long. As expected, the Medium version that doesn't reference endpoint adversaries (m2) performed poorly with the malware adversary.

4.2.5 Some messages may oversell e2e encryption

An important goal of our educational intervention is to avoid causing participants to believe that e2e encryption provides more security than it actually does. We found no significant results to this effect, but we do see some weak trends in the wrong direction. For example, Short messages that give an overview of e2e encryption but don't mention its weaknesses (s.conf, s.mech) show trends where participants may increase their belief that e2e encryption can protect metadata from app-company employees, the ISP, and the government. These trends can be seen in the extended paper (see Appendix D).

4.3 Summary of survey study results

Overall, we find that Long works better than Control, primarily for conveying information about the interception capability. Medium messages perform similarly to Long, and Short messages work reasonably well for relevant topics. We do not, however, see much improvement related to metadata weaknesses and the disadvantages of systems that do not use e2e encryption, primarily because participants seem to already have reasonably strong mental models for these topics.

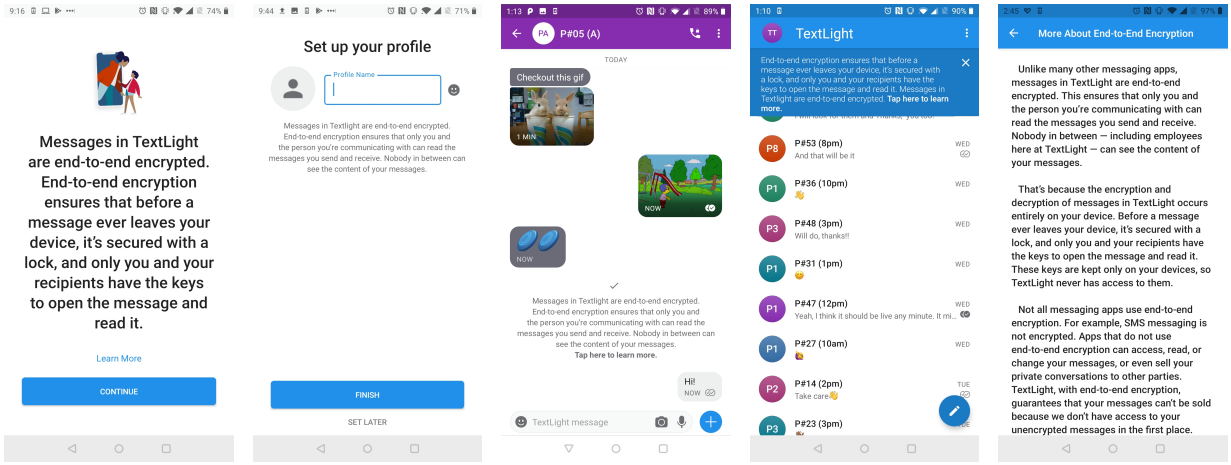
Overall, these results suggest optimism that integrating educational messages into app workflows may help to improve users' mental models. We therefore decided to conduct a second, more realistic, study to test these messages in context.

We opted to include all Short messages from the survey study in the follow-up. We hoped that including all messages would provide a reasonably complete view of e2e encryption and avoid overselling. Further, we hoped that including messages where many participants already had a correct mental model would reinforce when an existing model is correct.

5 App Study: Methods

Having found that the educational messages used in the survey study were reasonably useful in a controlled setting, we next designed a longitudinal app study to gauge their impact in a more realistic environment. Our participants (n=61) used a modified and rebranded version of the Signal messaging app [7] for Android (again called TextLight) for approximately three weeks. Half of the participants (n=32) used an experimental version of TextLight incorporating our Short and Long messages, while the other half (n=29) used a control version with no messages. We measure changes in mental models by comparing responses to a pre- and post-study communications privacy questionnaire similar to that used in the survey study. Participants in the experimental condition were invited to a post-study interview to provide more in-depth insight.

[7] https://signal.org/en/



(a) Splash (b) Profile (c) In-conversation (d) Conversation-list (e) Long

Figure 2: UI elements considered for delivery of the educational messages. All except (a) were used in the final study.

5.1 The TextLight App

Our participants interacted with TextLight, a modified version of the Signal Android client, branched from version 4.48.14 on October 8, 2019. We developed two versions of TextLight: an experimental condition that incorporated a variety of educational messages drawn from the survey study, and a control condition that contained no such messages[8]. By comparing these two versions, we can distinguish changes in mental models related to our educational intervention from any generic effects of using an app described as being e2e encrypted and answering encryption-related questions in our pre- and post-study questionnaires.

Educational messages As the Short messages were effective in the survey study, we decided to mainly incorporate Short messages into TextLight. We hypothesized that multiple Short messages could convey a broad overview of concepts without overwhelming the user with lengthy messages.

We made only minor modifications to the five Short messages from the survey study. We mentioned e2e encryption in general rather than TextLight specifically, and we reworded some messages slightly to differentiate them from each other and hopefully reduce habituation (see Appendix B).

For consistency with Signal's design language, we considered four existing UI elements that normally convey status information (e.g, a missed call) or prompt for action (e.g., to make the app the default messenger). These included a full-screen modal that appears on occasion when the app is opened(*splash screen*, Figure 2a); a message that appears when a user sets up their account (*profile*, Figure 2b); a grayed-out message that occasionally appears within a conversation thread and scrolls up as new messages are exchanged (*in-conversation*, Figure 2c); and a banner on top of the list of conversations (*conversation-list*, Figure 2d).

In addition, we made the Long message available as a full-page, scrollable message that could be accessed through the settings page of TextLight, by clicking on the conversation-list or in-conversation elements, or by selecting the "learn more" option from the splash-screen element. We refer to this as the *long* element (Figure 2e).

Our pilot participants reported being annoyed by the splash-screen element (several considered it a glitch); we therefore removed it before we launched the study.

Message display logic We set TextLight to show our short messages periodically, aiming to ensure users would see all messages while keeping low enough frequency to avoid annoyance. Messages are shown round-robin, in the following order: s.conf, s.mech, s.risk, s.meta, s.endpt.

The profile element is shown when the user opens the profile-setup or username-settings screens. An in-conversation message is shown each time the user initiates a conversation with a new recipient. This is similar to WhatsApp's current short notification that a new conversation is e2e encrypted. Messages are also shown probabilistically each time the user starts a new session with the app, based on the first screen visited. If the user starts on the list of conversations, they have a 20% chance of seeing the conversation-list message element; if they start within a conversation, they have a 20% chance of seeing an in-conversation message element. We constrained message frequency to ensure at least 8 hours between probabilistic messages but require a message if more than 43 hours have passed.

[8]Between the completion of the study and publication, the Signal architecture changed, causing TextLight to no longer work. As a result, the app was only partially evaluated by the USENIX Security artifact evaluation committee

Conversation-list messages persist until they are dismissed, but no more than one is displayed at a time. To ensure message rotation, we automatically dismiss any conversation-list message that has persisted for more than 10 hours, allowing a new message to (probabilistically) take its place.

We determined these frequency rules during piloting; based on pilot behavior, we expected approximately 0.75 messages per participant per day.

Other app modifications We made other minor changes from Signal to TextLight, including limiting unneeded features, instrumenting the app to measure our participants' interaction with it, streamlining the installation process for participants, and rebranding. We modified only the client app and used Signal's server-side infrastructure as is.

TextLight is designed only for use with our study, as our instrumentation is not compatible with the privacy goals of e2e encryption. To this end, we configured it to only work within our study setup, and we clearly marked it in the Google Play store as for a research study only.

We disabled unnecessary features that could create privacy risks for our participants or require connection to external services, including options to share Signal with friends, connect to a desktop client, use SMS, and any features that access the participant's local contacts or pre-existing text messages. We kept microphone, camera, and local storage permissions in case participants opted to share media messages.

We instrumented the app to measure how much time users spent on which pages, which UI elements they interacted with, and when they sent messages. We stored logs on our server, under participant pseudonyms we generated.

We also streamlined Signal's standard registration process for participants. After installing the app, participants only needed to enter a phone number (provided by us); we automated other verification and registration steps.

5.2 Study structure

The six stages our participants completed are detailed below[9].

Pre-screener and recruitment We again recruited from Prolific. This time, we pre-screened participants to rule out those with too much e2e encryption knowledge up front, rather than removing their responses after the fact. The study was advertised as "Messaging App Study" to avoid potential privacy-related selection bias. We invited participants to the study if:

- They resided in the U.S.;
- They had an Android 6.0 or above phone, in order to effectively use our TextLight app;
- They had never used Signal and would therefore not be biased by prior perceptions;
- They were e2e encryption novices: They disagreed or strongly disagreed that they could"describe what symmetric key encryption is," "describe what End-to-End Encryption is", and "describe a scenario where Diffie-Hellman key exchange is used." Since we expected the change in mental models to be more subtle than the survey study (due to the more realistic scenario), we selected for slightly less e2e encryption knowledge than previously; and
- They were willing to participate in a remote interview: Although only experimental participants were invited to interview, to avoid selection bias all participants were required to express willingness to interview.

Further, we collected IT background, gender, age, and education to ensure our sample was reasonably well balanced. We invited qualified participants to the main study (randomly assigned to either the control or experimental conditions).

Initial questionnaire Participants who were invited to the main study were first asked to consent to the entire study, then given a pre-intervention questionnaire. The questionnaire introduced TextLight and assured participants that the app was e2e encrypted. Next, we asked a slightly modified version of the communications privacy questionnaire described in Section 3.1. The main modification, based on piloting (Section 5.2), was to organize all not-e2e encryption questions together rather than distributing them among adversaries, to avoid confusion. We further distinguished these questions by tying them to a fictional non-e2e encrypted app we named MessageBright, to avoid any misinterpretation about TextLight. In addition, to reduce stress and discomfort, we remind users that there are no right or wrong answers.

Installation and tutorial Participants next viewed a tutorial on how to install and use TextLight. Because participants were recruited remotely and asynchronously, we aimed to make installation as seamless as possible.

Participants were instructed to install TextLight through the Google Play Store via a provided link. To minimize personal information collected, participants were provided a phone number controlled by the research team[10] to use for registration. We automated portions of Signal's registration confirmation process to minimize participant effort. Participants were shown an animation depicting a correct registration outcome.

Finally, via another animation, participants were instructed to start a new text-message conversation with a provided phone number. These numbers were operated by researchers; however, participants were not explicitly told who the numbers belonged to. To reduce bias, researchers did not know condition assignments; however, a few participants referenced educational messages unprompted during the daily conversations, and researchers learned three participants' conditions during tech support.

[9] Survey and interview protocols are given in the extended paper (see Appendix D)

[10] We obtained the numbers from Twilio.com, which also provides an API.

Application use After installation, participants were instructed to use the app to chat with the person they had started a conversation with, every day for 20 days. To count for compensation (detailed below), participants needed to send at least five messages per day, with the first and last messages at least 10 minutes apart. Participants were instructed not to share any information they considered private, and were notified that for study purposes the researchers would monitor some app uses, such as interaction with UI elements and how many messages were sent. (See Section 5.1 for instrumentation details.)

Researchers typically messaged participants each day for a brief conversation on generic (non-security or privacy) topics such as hobbies, daily news (non-political), sports, etc. Participants were occasionally instructed during conversation to initiate the next day's conversation themselves (three occurrences per participant) or to initiate a chat with a different number (a different researcher, two occurrences per participant). This forced participants to spend time on the screen that displays all conversations (Figure 2c), rather than only in a specific conversation (Figure 2c). These conversation patterns were designed to trigger (for the experimental group) our informational messages (see Section 5.1).

Exit questionnaire Twenty days after installation, we posted the exit questionnaire on Prolific and reminded the participants (through TextLight) to complete it. As in the survey study, we re-administered the communications privacy questionnaire (Section 5.2 used the exit questionnaire to obtain a post-intervention measurement of mental models by re-administering the communications privacy questionnaire (Section 5.2). We also asked questions about the participant's experience with the app, including the System Usability Scale [11], who they thought would use TextLight, whether they had noticed any bugs or glitches, and whether they had noticed "any informative messages or prompts." We also asked what participants thought was the purpose of the study.

Interviews We invited participants in the experimental condition who completed at least 18 of 20 conversation days to an exit interview. Interviews took on average less than 14 minutes. The goal of the interviews was to explore participants' mental models of e2e encryption and experiences with the app in more depth. We started with usability and general evaluation questions, including whether the app was easy to use and how it compared to other messenger apps.

We then asked questions related to our intervention messages, structured to test the participant's recall without reminding them of the messages. First, we asked if they remembered seeing any educational messages and where they were. We then showed screenshots, with the message text blurred out, as a prompt for recall. At that point, we asked participants if they could recall what the messages had said, whether it was the same message every time, how frequently they had seen the message, and whether they were interested in learning more about the content. Finally, we asked participants what they thought e2e encryption meant and what it would (not) protect against.

Compensation Participants were compensated $0.70 for completing the pre-screener, $2.00 for completing the initial questionnaire, $8.00 for installing the application and sending their first message, $1.50 per successful conversation day (as described in Section 5.2), $5.00 for completing the exit questionnaire, and $15.00 for completing an interview. Participants who dropped out before completing the exit questionnaire were not paid for conversation days. Average compensation for those who completed the entire study was $48.30 ($\sigma$ = $9.20); participants who started but didn't complete the study received on average $8.40 ($\sigma$ = $3.90).

Pilot testing We conducted (pre-COVID-19) three in-person pilots for the initial questionnaire and installation tutorial; two partially in-person pilots covering the entire study but with only 10 conversation days, and five fully online pilots covering the entire study but with only seven conversation days. In-person pilots were recruited from convenience samples; online ones were recruited from Prolific. Pilot testing helped us to refine study procedures, content and placement of educational messages, and questionnaire wording.

5.3 Data analysis

For the app study, we used a simplified version of the survey study analysis, with only one experimental group instead of eight. We again used differences between the pre- and post-intervention questionnaires as the main unit of analysis. We first confirmed that the capability groupings from the survey study (Section 3.3) still held. We then used two-tailed pairwise MWU tests to compare the control and experimental conditions for each adversary-capability set, reporting significance as well as effect size via location-shift estimates.

To check whether our educational messages reduce TextLight's usability, we compared SUS scores between the control and experimental conditions using the Mann-Whitney test for Equivalence (MWE) [68]. Unlike traditional hypothesis testing, the null hypothesis here that the two samples are different; if significant, they are likely to be drawn from the same distribution. We apply the stricter equivalence range suggested by Wellek [68].

Interviews were transcribed by a third-party service. Two researchers qualitatively coded the transcripts using an open-coding approach [14]. The two researchers established an initial codebook based on five randomly selected transcripts [51]. Then, the they independently coded two randomly selected interviews at a time to establish inter-rater reliability. After each batch, the researchers met to resolve differences and update the codebook. Once reliability was established on two interviews ($\alpha \geq .8$ [34]),researchers coded two more interviews (without resolving differences) to bring the set used

for reliability to ~20% of the interviews. As suggested by Campbell et al., one researcher unitized the interviews before coding in order keep the coded sections consistent [14]. We obtained a Krippendorff's α of .89.

As they were only a minor datapoint in our study, we collaboratively coded open-ended questions from the pre- and post-intervention questionnaires [39]. Note that there is some overlap between the interview and survey codebooks; we reuse already established codes when applicable.

5.4 Ethical considerations

This study was also approved by the University of Maryland IRB. We used standard ethics procedures, including obtaining consent before the pre-screener and again upon invitation to the main study; allowing participants to leave the study at any point with partial compensation; minimizing the collection of identifiable information; and keeping all potentially identifiable information on password-protected systems.

We considered pairing participants with each other for less mediated conversation, but decided not to in order to remove the potential for sending/receiving inappropriate messages. To further protect participants, we disabled certain Signal features to limit participants' exposure (Section 5.2) and asked participants not to share any private information during daily conversations. These decisions may limit ecological validity, but we considered them ethically necessary.

We collect demographic information such as age, ethnicity, and gender in order to report on the (un)representativeness of our samples (Sections 4.1 and 6.1). We offered "prefer not to answer" options for these questions.

5.5 Limitations

The app study was designed to address some of the ecological validity limitations of the survey study. However, other limitations typical for studies of this kind remain.

Our U.S.-based Prolific sample may not be sufficiently representative of the user base for messaging apps, as discussed in Section 3.5. Further, we limit the study to Android users. Possibly outdated research from 2014 suggests that Android users are more privacy sensitive, meaning they may be more interested in e2e encryption [45]. On the other hand, requiring participants to be willing to complete an interview may have selected for less privacy sensitivity. We attempt to mitigate this in part by limiting participation to users with little knowledge of e2e encryption.

While we attempted to approximate realistic use, texting two researchers as part of an experiment is not the same as using a messaging app with friends and family.

To protect participants, we instructed them not to share private information and alerted them to our instrumentation. This may reduce overall trust in e2e encryption and introduce unwanted bias. This may also reduce participants' investment in whether or not communications in TextLight are meaningfully private, which may limit interest in our educational messages. However, this was unavoidable to ethically protect participants. Further, our instrumentation is somewhat similar to the employee adversary and metadata capability we ask about. These issues apply to both the experimental and control conditions, enabling comparison.

When asked about the purpose of the study, participants generally assumed we were trying to test the features of a messaging application (n=41), and only three mentioned the educational messages. This suggests any demand effects would not be relevant to our research questions.

As mentioned in Section 3.5, non-parametric hypothesis tests have limited power, meaning subtle shifts in mental models may not manifest in test results. A-priori power analysis indicated 30 participants per group would be enough to detect large effects (Cohen's $d = 0.8$ [15]) with 80% power but not enough to meet the same standards of the survey study (Long vs. Control). For that, we would have to recruit 30 more participants per group which was not feasible for our costly experimental setup (time-consuming interaction with participants). Instead, we recruit people less knowledgeable about e2e encryption (see 5.2) for more obvious mental model changes and gather extensive qualitative data (interviews, open-ended survey questions) to add depth to our results.

6 App Study: Results

We next detail the results of the app study.

6.1 Participants

We received 261 prescreening responses, of which 89 qualified and 84 were invited to the main study. We invited in batches, stopping once we had at least 65 participants actively using TextLight. (We aimed for about 60 valid participants after expected dropouts.) Sixty-eight participants started the main study. We disqualified five participants for missing too many conversation days (despite reminders) or uninstalling TextLight. In total, 61 participants (32 experimental, 29 control) completed the exit questionnaire. We invited 23 of the 32 experimental participants for an interview; 19 agreed to participate. Data was collected in April and May 2020.

Table 2 shows demographics of our app study participants. which are similar to the survey study.

6.2 Using TextLight

Most participants used the app in line with our goals. Participants completed an average of 18.5 conversation days (σ = 3.3) with an average of 156.0 minutes (σ=135.1) of screen time in TextLight over the duration. Participants spent an average of 139.2 minutes (σ = 122.6) in the conversation

screen and sent on average 138.2 (σ = 44.9) messages, more than the required 100 over 20 days.

To investigate whether the educational messages interfered with the usability of the app, we compared the SUS scores of the experimental and control group using the MWE test. We found no difference in usability between them (p = 0.026).

Our interviewees (experimental condition only) generally found TextLight easy to use (n=19), professionally designed (n=12), and similar to other messaging apps (n=11). These responses may be influenced by demand effects, as participants generally assumed we were testing a new app we had developed, and may have wanted to say nice things. Nonetheless, we believe these responses suggest TextLight was sufficiently comparable to a real app to meet our ecological validity goals.

Only one participant noted e2e encryption when comparing TextLight to other messaging apps.When asked about features that stood out, five mentioned our educational messages and two mentioned security features without referencing the messages directly. When asked in the exit questionnaire about who might want to use TextLight, 39 of 61 (23 experimental, 16 control) answers mentioned privacy or security. Of these, 11 (6 experimental, 5 control) mentioned the need for security and privacy for professionals such as "people who conduct private business" (P56, experimental) or "doctors with patients' health conditions" (P11, experimental). A large minority (n=26, 10 experimental and 16 control) mentioned general-purpose users unrelated to privacy or security (e.g., from the experimental group P32 said, "All the regular people that communicate through text messaging").

6.3 Encountering educational messages

Experimental participants saw on average 19.4 e2e encryption messages during the study. Of these, 10.7 were in-conversation messages, 6.5 were conversation-list, 1.4 were profile, and 0.9 were long. Long messages, which required participants to take explicit action, were only seen by 18 participants. Within this group, Long was opened on average 1.6 times and was displayed on average for 19.0 seconds over the duration of the study (σ = 26.71). All 32 participants saw all three other kinds of messages. All five Short message versions were viewed approximately the same number of times ($\sim$ 3.7).

Most remembered the educational messages In the exit questionnaire, most experimental participants said they did see "informative messages or prompts" (n=23) while others said they didn't see (n=3) or didn't remember (n=5) the messages. Most (n=21) remembered that the messages were about e2e encryption; however, two experimental participants who claimed to remember described unrelated messages.

The interviews provide more hints about the effectiveness of the e2e encryption messages. Thirteen of 19 interviewees recalled the messages without prompting; of these, seven described the conversation-list messages and eight described the in-conversation messages (some overlap). After being shown blurred screenshots, only four remembered the profile message, 17 remembered in-conversation messages, and 13 recalled conversation-list messages. Participants who remembered the messages (n=17) generally said they saw them either every day (n=7) or every second day (n=6).

However, most paid them little attention During the interview, four participants explicitly said they ignored the educational messages. Another seven gave responses indicating habituation. For instance, P15 said, "I don't think that I really thought to read it because I assumed that it was some type of generic welcoming message or something probably."

When asked in the interview if they were intrigued by the messages or wanted to learn more, seven said they weren't interested and six said they were (although this may be exaggerated by demand effects). Only two participants said they clicked on the short messages in order to "learn more"; however, our logs show that 18 experimental participants did click on a Short message and access the Long message. Three participants also accessed Long through the settings menu.

When asked to recall whether the messages varied, most participants said there was only one version (n=8) or that they did not recall (n=6). In fact, there were five messages.

We also asked the 17 participants who recalled seeing the messages about their content. Six mentioned e2e encryption but could not give further specifics. A few mentioned specific concepts we aimed to convey: weakness against metadata (n=3), that no one can read sent messages (n=2), or that only endpoints can read messages (n=2). Others mistakenly reported that the messages were about how to use TextLight or simply said they did not remember.

Taken together, these comments suggest that participants noticed the messages existed but did not examine them carefully, as might be expected in a real-world scenario.

6.4 Mental models of e2e encryption

Unfortunately, we found no statistical evidence, in comparing the experimental and control conditions, that our educational messages improved mental models. Our interviews with experimental participants shed light on why the messages were less effective than we hoped.

No significant improvements in the questionnaire We found only one significant difference in perceptions of adversary capability (Table 4). Experimental participants were somewhat less likely to believe app-company employees could observe metadata (p=0.03; location shift estimate -1), which is a change in the wrong direction. This fits our survey-study observation that short messages can sometimes oversell the benefits of e2e encryption; our hope that rotating through the messages would mitigate this issue was not borne out.

A closer look at effect sizes shows that the adversary/capability pairs with the largest effect sizes in the survey study (Interception capabilities of Employee and Government

	Employee		ISP		Server Hacker		Government		Unlocked		Malware	
	S	A	S	A	S	A	S	A	S	A	S	A
Interception	-1.00*	-0.67	-0.67*	0.00	0.00	0.00	-1.00*	-0.33	0.00	0.00	0.67*	0.00
Metadata	0.00	-1.00*	0.00*	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Not-E2EE	n/a	0.00	0.00	1.00	n/a	0.00	n/a	0.00	n/a	0.00	n/a	0.00

Table 4: Location-shift estimates for MWU test results. The left column for each adversary-capability compares Control to Long in the survey study; the right column compares Control to Experimental in the app study. * indicates a statistically significant result ($p < 0.05$).

adversaries in Table 4) also appear to have negative effects (the correct direction) in the app study. Although encouraging, it is not clear if this is a real effect that is too small for our experiment to confirm or just noise. Further research would be needed to validate these trends.

Similar to the survey study, we observe ceiling effects with not-e2e encryption and metadata capabilities with most adversaries[11].

Definitions of e2e encryption are high-level but mainly correct We asked interview participants to define e2e encryption. The most common response conveyed that only the sender and intended recipient could view the content of the message (n=8), as exemplified by P13: "A message that you send out is encrypted and the only person who can unencrypt it to read it would be the receiver of the message." Other definitions differed slightly by emphasizing who could *not* view (n=5), alter (n=5), or otherwise intercept (n=4) the message. Three participants used the key-lock metaphor we described in the s.mech and Long messages, and three mentioned the metadata weakness detailed in s.meta. Only two participants said they were not sure what e2e encryption meant.

Recognition of protection from non-endpoint threats When asked to explain what e2e encryption protects against, about half of interview participants (n=10) generally described it as effective against non-endpoint adversaries. P23, for example, said, "Probably anyone who would interrupt or interfere in between the messaging, in between where you sent it and someone else received it."

Participants also frequently mentioned protection from adversaries highlighted in the educational messages and/or the communications privacy questionnaire. Four mentioned a server hacker, four mentioned the ISP, and three mentioned the government adversary. Three mentioned ambiguous adversaries such as "hackers," and three incorrectly suggested e2e encryption would protect against malware.

e2e encryption weaknesses were less clear When asked what e2e encryption does not protect against, participants again mentioned adversaries we described in the educational messages and questionnaires. Most mentioned an unlocked phone (n=14); in both studies, we found that participants largely started with a correct mental model for this before our intervention. A few participants mentioned the government (n=4), an app company employee (n=2), or a server hacker (n=1). Three specifically noted that e2e encryption could not protect against all "hackers." In total, nine of 19 participants gave answers that at least in part contradicted the principles we attempted to convey in the educational messages. As one example, P11 said, "The company essentially has access to it. They don't necessarily look at it, but if the proper legal methods are observed, there is a chance that someone else might be able to see it, for instance, the government."

[11] These are illustrated in the extended paper (see Appendix D).

7 Discussion

Our educational messages were effective in isolation, but when embedded into app workflows they did not show statistically significant effects. This is likely related to the fact that although most participants noticed the messages, many ignored their contents, possibly out of habituation to informational messages generally. The difference may also reflect short-term recall in the survey study compared to longer-term recall in the app study. Overall, this suggests messages like ours may need to be somewhat more intrusive to be useful.

Educational intervention works, to an extent In the survey study, our educational interventions worked reasonably well, with minimal unintended consequences. In line with our prior work [9], participants easily grasped core principles related to confidentiality (measured via our interception capability). To some extent, participants gained understanding about metadata weaknesses. However, we did find some evidence in the app study that our intervention may have oversold the capabilities of e2e encryption with respect to metadata.

We also found evidence that many participants already possessed strong mental models with respect to risks of not-e2e encryption communications and risks of physical access at endpoints. These findings reflect somewhat more knowledge than was observed in prior work [2, 3, 17] — this may reflect differences in study populations, or that users are learning as they gain exposure to e2e encrypted apps over time. We argue that where participants have correct models like these, educational interventions should reinforce them.

Interventions may need to be more intrusive Unfortunately, we were unable to replicate the successes of the survey

study in a more realistic in-workflow context. We mainly attribute this difference to the messages failing to attract sufficient user attention in this more realistic setting.

However, quantitative and qualitative results suggest that participants did not find our interventions intrusive or unusable; thus, there may be room to make such interventions more noticeable without triggering an undue amount of user annoyance. As one example, we decided during pilot testing to remove the splash-screen message element that participants found somewhat disruptive; in hindsight, we hypothesize that this might have struck a better balance between usability and noticeability. Other modifications could include making the educational messages bigger or bolder, highlighting key phrases, or using graphics to make them more eye-catching. Future work should explore whether changes like these can achieve better results without significant harm to usability.

Experimental setup The discrepancy between the two studies could also be attributed to the differing experimental setups. The survey study involved one intervention with questions, on average, less than 10 minutes later. The app study involved 20 days of participation with interventions every 1-2 days. Thus, we might have measured short-term recall of educational material with the survey study, vs. longer-term impact on mental models with the app study. On the other hand, prior work provides some evidence that security nudging surveys can have longer term impact [55]. Additional controlled experiments would be needed to know to what extent our survey study had lasting impact on mental models.

Other kinds of interventions Our results also underscore that in-workflow messages are only one way to influence mental models of secure communication. As in our prior work [8, 9], when our participants were focused on our educational content, they did learn functional information. While it is not realistic to expect most users to seek out training on secure communication, this result bolsters the importance of making well-designed educational materials available to those who do seek them out. Organizations like EFF and the Library Freedom Project [20, 42] have developed several such materials; future work should consider evaluating where they succeed and whether improvements can be made.

Further, there is increasing emphasis on teaching everyday privacy and security concepts in elementary and secondary schools [35]. Including functional models of secure communication in these curricula could help these students, as they grow up, to make appropriate choices about their communications mechanisms in an increasingly networked world.

8 Conclusion

In this work, we created educational messages to improve functional mental models of e2e encryption and evaluated them in both a controlled and a more realistic setting. We find that conveying functional mental models of e2e encryption is possible in isolation, but we hypothesize in-app nudging may require more intrusiveness to be effective; more experiments are needed.

9 Acknowledgements

We thank our participants. This material is based upon work supported by the United States Air Force and DARPA under Contract No FA8750-16-C-0022. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force and DARPA.

References

[1] Whatsapp hack: Is any app or computer truly secure? *BBC*, 2019.

[2] Ruba Abu-Salma, Elissa M Redmiles, Blase Ur, and Miranda Wei. Exploring User Mental Models of End-to-End Encrypted Communication Tools. In *FOCI*, 2018.

[3] Ruba Abu-Salma, M Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. Obstacles to the adoption of secure communication tools. In *IEEE S&P*, 2017.

[4] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Comput. Surv.*, 50(3), August 2017.

[5] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your Location Has Been Shared 5,398 Times! A Field Study on Mobile App Privacy Nudging. In *CHI*, 2015.

[6] Apple Inc. *Privacy - Approach to Privacy*. (Last accessed on Sep. 2019).

[7] Farzaneh Asgharpour, Debin Liu, and L. Jean Camp. Mental Models of Security Risks. In *USEC*, 2007.

[8] Wei Bai, Moses Namara, Yichen Qian, Patrick Gage Kelley, Michelle L Mazurek, and Doowon Kim. An Inconvenient Trust: User Attitudes Toward Security and Usability Tradeoffs for Key-Directory Encryption Systems. In *SOUPS*, 2016.

[9] Wei Bai, Michael Pearson, Patrick Gage Kelley, and Michelle L Mazurek. Improving Non-Experts' Understanding of End-to-End Encryption: An Exploratory Study. In *EuroUSEC*, 2020.

[10] James Ball. Nsa collects millions of text messages daily in 'untargeted' global sweep. *The Guardian*, 2014.

[11] John Brooke. SUS: a Quick and Dirty Usability Scale. In *Usability Evaluation in Industry*. CRC press, 1996.

[12] José Carlos Brustoloni and Ricardo Villamarín-Salomón. Improving Security Decisions with Polymorphic and Audited Dialogs. In *SOUPS*, 2007.

[13] L. J. Camp. Mental Models of Privacy and Security. *IEEE Technology and Society Magazine*, 28(3):37–46, Fall 2009.

[14] John L Campbell, Charles Quincy, Jordan Osserman, and Ove K Pedersen. Coding In-Depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement. *Sociological Methods & Research*, 42(3):294–320, 2013.

[15] Jacob Cohen. A power primer. *Psychological bulletin*, 112(1):155, 1992.

[16] Joseph Cox. China Is Forcing Tourists to Install Text-Stealing Malware at its Border. *Vice*, 2019.

[17] S. Dechand, A. Naiakshina, A. Danilova, and M. Smith. In Encryption We Don't Trust: The Effect of End-to-End Encryption to the Masses on User Perception. In *EuroS&P*, 2019.

[18] A Demjaha, JM Spring, I. Becker, S Parkin, and MA Sasse. Metaphors Considered Harmful? An Exploratory Study of the Effectiveness of Functional Metaphors for End-to-End Encryption. In *USEC*, 2018.

[19] Andrea A. diSessa. Models of Computation. In Donald A. Norman and Stephen W. Draper, editors, *User Centered System Design: New Perspectives on Human-Computer Interaction*, pages 201–218. Lawrence Erlbaum Associates, 1986.

[20] Electronic Frontier Foundation. Communicating with others. https://ssd.eff.org/en/module/communicating-others.

[21] Antonio M. Espinoza, William J. Tolley, Jedidiah R. Crandall, Masashi Crete-Nishihata, and Andrew Hilts. Alice and Bob, Who the FOCI Are They?: Analysis of End-to-End Encryption in the LINE Messaging Application. In *FOCI*, 2017.

[22] Facebook. Secret conversations. https://www.facebook.com/help/messenger-app/1084673321594605.

[23] Facebook. Whatsapp security. https://www.whatsapp.com/security/.

[24] Sascha Fahl, Marian Harbach, Thomas Muders, Matthew Smith, and Uwe Sander. Helping Johnny 2.0 to Encrypt His Facebook Conversations. In *SOUPS*, 2012.

[25] Simson L. Garfinkel and Robert C. Miller. Johnny 2: A User Test of Key Continuity Management with S/MIME and Outlook Express. In *SOUPS*, 2005.

[26] Shirley Gaw, Edward W. Felten, and Patricia Fernandez-Kelly. Secrecy, Flagging, and Paranoia: Adoption Criteria in Encrypted Email. In *CHI*, 2006.

[27] Nina Gerber, Verena Zimmermann, Birgit Henhapl, Sinem Emeröz, and Melanie Volkamer. Finally Johnny Can Encrypt: But Does This Make Him Feel More Secure? In *ARES*, 2018.

[28] Joseph A Gliem and Rosemary R Gliem. Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales. Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, 2003.

[29] Eszter Hargittai and Yuli Patrick Hsieh. Succinct Survey Measures of Web-Use Skills. *Social Science Computer Review*, 30(1):95–107, 2012.

[30] Kashmir Hill. 'God View': Uber allegedly stalked users for party-goers' viewing pleasure. *Forbes*, 2014.

[31] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric Statistical Methods*, volume 751. John Wiley & Sons, 2013.

[32] Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, pages 65–70, 1979.

[33] Patrick Gage Kelley, Sunny Consolvo, Lorrie Faith Cranor, Jaeyeon Jung, Norman Sadeh, and David Wetherall. A conundrum of permissions: installing applications on an android smartphone. In *Financial Crypto*, 2012.

[34] Klaus Krippendorff. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human communication research*, 30(3):411–433, 2004.

[35] Priya C. Kumar, Marshini Chetty, Tamara L. Clegg, and Jessica Vitak. Privacy and Security Considerations For Digital Technology Use in Elementary Schools. In *CHI*, 2019.

[36] Alexander De Luca, Sauvik Das, Martin Ortlieb, Iulia Ion, and Ben Laurie. Expert and Non-Expert Attitudes towards (Secure) Instant Messaging. In *SOUPS*, 2016.

[37] Hiroaki Masaki, Kengo Shibata, Shui Hoshino, Takahiro Ishihama, Nagayuki Saito, and Koji Yatani. Exploring Nudge Designs to Help Adolescent SNS Users Avoid Privacy and Safety Threats. In *CHI*, 2020.

[38] Arunesh Mathur, Josefine Engel, Sonam Sobti, Victoria Chang, and Marshini Chetty. "They Keep Coming Back Like Zombies": Improving Software Updating Interfaces. In *SOUPS*, 2016.

[39] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Quidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[40] Alena Naiakshina, Anastasia Danilova, Sergej Dechand, Kat Krol, M Angela Sasse, and Matthew Smith. Poster: Mental Models-User Understanding of Messaging and Encryption. In *EuroUSEC*, 2016.

[41] Eyal Peer, Serge Egelman, Marian Harbach, Nathan Malkin, Arunesh Mathur, and Alisa Frik. Nudge Me Right: Personalizing Online Security Nudges to People's Decision-Making Styles. *Computers in Human Behavior*, 109:106347, 2020.

[42] Library Freedom Project. Library freedom resouces. https://libraryfreedom.org/index.php/resources/.

[43] Rakuten. Viber security. https://www.viber.com/security/.

[44] Rakuten. Viber: Support portal. https://support.viber.com/customer/en/portal/articles/2017401-viber-accounts-security-and-encryption.

[45] Lena Reinfelder, Zinaida Benenson, and Freya Gassmann. Differences Between Android and iPhone Users in their Security and Privacy Awareness. In *TrustBus*, 2014.

[46] Anna L Rowe and Nancy J Cooke. Measuring mental models: Choosing the right tools for the job. *Human resource development quarterly*, 6(3):243–255, 1995.

[47] Scott Ruoti, Jeff Andersen, Scott Heidbrink, Mark O'Neill, Elham Vaziripour, Justin Wu, Daniel Zappala, and Kent Seamons. We're on the Same Page: A Usability Study of Secure Email Using Pairs of Novice Users. In *CHI*, 2016.

[48] Scott Ruoti, Jeff Andersen, Tyler Monson, Daniel Zappala, and Kent Seamons. A Comparative Usability Study of Key Management in Secure Email. In *SOUPS*, 2018.

[49] Scott Ruoti et al. A Usability Study of Four Secure Email Tools Using Paired Participants. *ACM Transactions on Privacy and Security (TOPS)*, 22(2):13, 2019.

[50] Scott Ruoti, Nathan Kim, Ben Burgon, Timothy Van Der Horst, and Kent Seamons. Confused Johnny: When Automatic Encryption Leads to Confusion and Mistakes. In *SOUPS*, 2013.

[51] Johnny Saldaña. *The Coding Manual for Qualitative Researchers*. Sage, 2015.

[52] Wickr Security. *Your Conversations and Data are Private by Design*. https://wickr.com/security/.

[53] Steve Sheng, Levi Broderick, Colleen Alison Koranda, and Jeremy J Hyland. Why Johnny Still Can't Encrypt: Evaluating the Usability of Email Encryption Software. In *SOUPS*, 2006.

[54] Signal Foundation. Signal terms and privacy policy. https://signal.org/legal/.

[55] Peter Story, Daniel Smullen, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. From intent to action: Nudging users towards secure mobile payments. In *SOUPS*, 2020.

[56] Joshua Tan, Lujo Bauer, Joseph Bonneau, Lorrie Faith Cranor, Jeremy Thomas, and Blase Ur. Can Unicorns Help Users Compare Crypto Key Fingerprints? In *CHI*, 2017.

[57] Telegram F.A.Q. Secret chats. https://telegram.org/faq#secret-chats.

[58] Telegram F.A.Q. What is telegram? https://telegram.org/faq#q-what-is-telegram-what-do-i-do-here.

[59] Threema. *Cryptography Whitepaper*, 2019. https://threema.ch/press-files/cryptography_whitepaper.pdf.

[60] Wenley Tong, Sebastian Gold, Samuel Gichohi, Mihai Roman, and Jonathan Frankle. Why King George III Can Encrypt. *Freedom to Tinker*, 2014.

[61] Blase Ur, Felicia Alfieri, Maung Aung, Lujo Bauer, Nicolas Christin, Jessica Colnago, Lorrie Cranor, Harold Dixon, Pardis Emami Naeini, Hana Habib, Noah Johnson, and William Melicher. Design and Evaluation of a Data-Driven Password Meter. In *CHI*, 2017.

[62] Elham Vaziripour, Devon Howard, Jake Tyler, Mark O'Neill, Justin Wu, Kent Seamons, and Daniel Zappala. I Don't Even Have to Bother Them! Using Social Media to Automate the Authentication Ceremony in Secure Messaging. In *CHI*, 2019.

[63] Elham Vaziripour, Justin Wu, Mark O'Neill, Daniel Metro, Josh Cockrell, Timothy Moffett, Jordan Whitehead, Nick Bonner, Kent Seamons, and Daniel Zappala. Action Needed! Helping Users Find and Complete the Authentication Ceremony in Signal. In *SOUPS*, 2018.

[64] Elham Vaziripour, Justin Wu, Mark O'Neill, Jordan Whitehead, Scott Heidbrink, Kent Seamons, and Daniel Zappala. Is that you, Alice? A Usability Study of the Authentication Ceremony of Secure Messaging Applications. In *SOUPS*, 2017.

[65] Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. A Field Trial of Privacy Nudges for Facebook. In *CHI*, 2014.

[66] Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. Privacy Nudges for Social Media: An Exploratory Facebook Study. In *WWW*, 2013.

[67] Rick Wash. Folk Models of Home Computer Security. In *SOUPS*, 2010.

[68] Stefan Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press, 2010.

[69] WhatsApp. *Two Billion Users – Connecting the World Privately*. https://blog.whatsapp.com/two-billion-users-connecting-the-world-privately?_fb_noscript=1.

[70] WhatsApp. How we work with facebook companies, 2020. https://faq.whatsapp.com/general/security-and-privacy/how-we-work-with-the-facebook-companies?eea=1&_fb_noscript=1.

[71] WhatsApp FAQ. *End-to-End Encryption*. (Last accessed on Sep. 2019).

[72] Justin Wu, Cyrus Gattrell, Devon Howard, Jake Tyler, Elham Vaziripour, Daniel Zappala, and Kent Seamons. "Something isn't secure, but I'm not sure how that translates into a problem": Promoting Autonomy by Designing for Understanding in Signal. In *SOUPS*, 2019.

[73] Justin Wu and Daniel Zappala. When is a Tree Really a Truck? Exploring Mental Models of Encryption. In *SOUPS*, 2018.

Appendix

A Messages used in the survey study

A.1 Long message

Unlike many other messaging apps, messages in *TextLight* are end-to-end encrypted. This ensures that **only you and the person you're communicating with can read the messages you send and receive. Nobody in between – including employees here at *TextLight* – can see the content of your messages.**

That's because the encryption and decryption of messages in *TextLight* occurs entirely on your device. Before a message ever leaves your device, it's secured with a lock, and only you and your recipients have the keys to open the message and read it. **These keys are kept only on your devices, so *TextLight* never has access to them.**

Not all messaging apps use end-to-end encryption. For example, SMS messaging is not encrypted. **Apps that do**

not use end-to-end encryption can access, read, or change your messages, or even sell your private conversations to other parties. *TextLight*, with end-to-end encryption, guarantees that your messages can't be sold because we don't have access to your unencrypted messages in the first place. With end-to-end encrypted messaging, privacy isn't just a promise; it's mathematically ensured.

It's important to note that *TextLight* and end-to-end encryption cannot protect against all possible privacy threats. **Parties such as *TextLight* and your internet service provider (Verizon, T-mobile, etc.) may be able to tell that you are exchanging messages with someone,** even if they can't see what the messages say. Moreover, **end-to-end encryption cannot protect you from someone who gets their hands on your unlocked phone, or from a hacker who has successfully installed malicious software.** *TextLight* also does not prevent your correspondent from publishing the messages you have exchanged.

A.2 Medium messages

m1: No confidentiality Messages in TextLight are end-to-end encrypted. Before a message ever leaves your device, it's secured with a lock, and only you and your recipients have the keys to open the message and read it. These keys are kept only on your devices, so TextLight never has access to them.

Apps that do not use end-to-end encryption can read, change your messages, or even sell your private conversations to other parties. With end-to-end encrypted messaging, privacy isn't just a promise; it's mathematically ensured.

It's important to note that TextLight and end-to-end encryption cannot protect against all possible privacy threats. Parties such as TextLight and your internet service provider (Verizon, T-mobile, etc.) may be able to tell that you are exchanging messages with someone, even if they can't see what the messages say. Moreover, end-to-end encryption cannot protect you from someone who gets their hands on your unlocked phone, or from a hacker who has successfully installed malicious software.

m2: No endpoint weakness Messages in TextLight are end-to-end encrypted. This ensures that only you and the person you're communicating with can read the messages you send and receive.

That's because before a message ever leaves your device, it's secured with a lock, and only you and your recipients have the keys to open the message and read it. These keys are kept only on your devices, so TextLight never has access to them.

Apps that do not use end-to-end encryption can read, change your messages, or even sell your private conversations to other parties. With end-to-end encrypted messaging, privacy isn't just a promise; it's mathematically ensured.

It's important to note that TextLight and end-to-end encryption cannot protect against all possible privacy threats. Parties such as TextLight and your internet service provider (Verizon, T-mobile, etc.) may be able to tell that you are exchanging messages with someone, even if they can't see what the messages say.

A.3 Short messages

Confidentiality (s.conf): Messages in TextLight are end-to-end encrypted. This ensures that only you and the person you're communicating with can read the messages you send and receive. Nobody in between can see the content of your messages.

Metadata weakness (s.meta): Messages in TextLight are end-to-end encrypted. However, parties such as TextLight and your internet service provider (Verizon, T-mobile, etc.) may be able to tell that you are exchanging messages with someone, even if they can't see what the messages say.

Endpoint weakness (s.endpt): Messages in TextLight are end-to-end encrypted. However, end-to-end encryption cannot protect you from someone who gets their hands on your unlocked phone, or from a hacker who has successfully installed malicious software.

Mechanism (s.mech): Messages in TextLight are end-to-end encrypted. Before a message ever leaves your device, it's secured with a lock, and only you and your recipients have the keys to open the message and read it.

Risks (s.risk): Messages in TextLight are end-to-end encrypted however, not all messaging apps use end-to-end encryption. Apps that do not use end-to-end encryption can read, change, or even sell your messages to other parties.

A.4 Control message

TextLight is a messaging app with a focus on speed and accuracy, it's super-fast, simple and free. You can use TextLight on all your devices at the same time – your messages sync seamlessly across any number of your phones, tablets or computers.

With TextLight, you can send messages, photos, videos and files of any type (doc, zip, mp3, etc), as well as create groups for up to 200,000 people or channels for broadcasting to unlimited audiences. You can write to your phone contacts and find people by their usernames. As a result, TextLight is like SMS and email combined – and can take care of all your personal or business messaging needs. In addition to this, we support voice calls.

B Messages used in the app study

B.1 Short messages

- Messages in TextLight are end-to-end encrypted. End-to-end encryption ensures that only you and the person you're communicating with can read the messages you send and receive. Nobody in between can see the content of your messages.
- Even though messages in TextLight are end-to-end encrypted, parties such as TextLight and your internet/mobile service provider (Verizon, AT&T, etc.) may still be able to tell that you are exchanging messages with someone, even if they can't see what the messages say.
- We use end-to-end encryption in TextLight to keep your messages safe, but end-to-end encryption cannot protect you from someone who gets their hands on your unlocked phone, or from a hacker who has successfully installed malicious software on your phone.
- End-to-end encryption ensures that before a message ever leaves your device, it's secured with a lock, and only you and your recipients have the keys to open the message and read it. Messages in TextLight are end-to-end encrypted.
- Not all messaging apps use end-to-end encryption. Apps that do not use end-to-end encryption can read, change, or even sell your messages to others. That's why we always use end-to-end encryption in TextLight.

B.2 Long message

Unlike many other messaging apps, messages in TextLight are end-to-end encrypted. End-to-end encryption ensures that only you and the person you're communicating with can read the messages you send and receive. Nobody in between – including employees here at TextLight – can see the content of your messages.

That's because with end-to-end encryption, the encryption and decryption of messages in TextLight occurs entirely on your device. Before a message ever leaves your device, it's secured with a lock, and only you and your recipients have the keys to open the message and read it. These keys are kept only on your devices, so TextLight never has access to them.

Not all messaging apps use end-to-end encryption. For example, SMS messaging is not encrypted. Apps that do not use end-to-end encryption can access, read, or change your messages, or even sell your private conversations to other parties. TextLight, with end-to-end encryption, guarantees that your messages can't be sold because we don't have access to your unencrypted messages in the first place. With end-to-end encrypted messaging, privacy isn't just a promise; it's mathematically ensured.

We use end-to-end encryption in TextLight to keep your messages safe, but end-to-end encryption cannot protect against all possible privacy threats. Parties such as TextLight and your internet/mobile service provider (Verizon, AT&T, etc.) may be able to tell that you are exchanging messages with someone, even if they can't see what the messages say. Moreover, end-to-end encryption cannot protect you from someone who gets their hands on your unlocked phone, or from a hacker who has successfully installed malicious software on your phone. TextLight also does not prevent your correspondent from publishing the messages you have exchanged.

C Mapping of educational messages to findings of our previous work [9]

A mapping between suggestions in our previous work and educational messages in this work are given in this Appendix. All short messages mentioned below are also included in the Long message.

- **Confidentiality:** Previously we had suggested that users find confidentiality the most important aspect of e2e encryption. We emphasize this point in s.conf, s.mech.
- **Risks:** We directly communicate the risks of not using e2e encrypted systems with s.risk. Our previous work noted that that conveying the risks of not using e2e encryption was important and surprising to multiple users.
- **Mechanism:** We dedicate a short message to the inner workings of e2e encryption (s.mech). As suggested previously, we use a brief analogy and emphasize that this mechanism ensures confidentiality. This aims to strike a balance between users who want to learn more about the technical aspects of e2e encryption, and users who are confused by it.
- **Endpoint and metadata weaknesses:** We had previously suggested that it is important to mention weaknesses of e2e encryption, we achieve this with s.meta and s.endpt. s.endpt emphasizes that the endpoints aren't protected by e2e encryption, and s.meta conveys that e2e encryption by itself does not protect metadata. Although metadata weakness wasn't emphasized in our prior work, we think it's an important limitation of e2e encryption and therefore include it.

In addition, we follow other recommendations such as avoiding integrity and authenticity topics, employing risk communication methods, and integrating the messages in regular communication workflows.

D Extended Appendices

An extended version of the paper with instrumentation and additional figures based on results can be found at: https://github.com/SP2-MC2/e2ee.