A Usability Study of Four Secure Email Tools Using Paired Participants

SCOTT RUOTI, University of Tennessee, Knoxville
JEFF ANDERSEN, LUKE DICKINSON, SCOTT HEIDBRINK, TYLER MONSON, MARK O'NEILL, KEN REESE, BRAD SPENDLOVE, ELHAM VAZIRIPOUR, JUSTIN WU, DANIEL ZAPPALA, and KENT SEAMONS, Brigham Young University

Secure email is increasingly being touted as usable by novice users, with a push for adoption based on recent concerns about government surveillance. To determine whether secure email is ready for grassroots adoption, we employ a laboratory user study that recruits pairs of novice users to install and use several of the latest systems to exchange secure messages. We present both quantitative and qualitative results from 28 pairs of novices as they use Private WebMail (Pwm), Tutanota, and Virtru and 10 pairs of novices as they use Mailvelope. Participants report being more at ease with this type of study and better able to cope with mistakes since both participants are "on the same page." We find that users prefer integrated solutions over depot-based solutions and that tutorials are important in helping first-time users. Finally, our results demonstrate that Pretty Good Privacy using manual key management is still unusable for novice users, with 9 of 10 participant pairs failing to complete the study.

CCS Concepts: • **Security and privacy** → *Domain-specific security and privacy architectures*; **Usability in security and privacy**; • **Human-centered computing** → **User studies**; **Empirical studies in HCI**; *Laboratory experiments*;

Additional Key Words and Phrases: Grassroots adoption, paired participants, PGP, secure email

ACM Reference format:
Scott Ruoti, Jeff Andersen, Luke Dickinson, Scott Heidbrink, Tyler Monson, Mark O'Neill, Ken Reese, Brad Spendlove, Elham Vaziripour, Justin Wu, Daniel Zappala, and Kent Seamons. 2019. A Usability Study of Four Secure Email Tools Using Paired Participants. *ACM Trans. Priv. Secur.* 22, 2, Article 13 (April 2019), 33 pages.
https://doi.org/10.1145/3313761

1 INTRODUCTION

In recent years, there has been an increase in the promotion of secure email, with tools such as Tutanota, Virtru, Mailvelope, ProtonMail, StartMail, Hushmail and others being pitched for every-

This work was supported in part by the National Science Foundation under grant CNS-1528022. This work was completed while S. Ruoti, M. O'Neill, and S. Heidbrink were interns for Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.
Authors' addresses: S. Ruoti, Min H. Kao Department of Electrical Engineering & Computer Science, 1520 Middle Dr, Knoxville, TN 37996-2250; email: ruoti@utk.edu; J. Andersen, L. Dickinson, S. Heidbrink, T. Monson, M. O'Neill, K. Reese, B. Spendlove, E. Vaziripour, J. Wu, D. Zappala, and K. Seamons, Computer Science Department, Brigham Young University, 3361 TMCB PO Box 26576, Provo, UT 84602-6576; emails: luke@isrl.byu.edu, sheidbri@byu.edu, monson@isrl.byu.edu, {brad.spendlove, elhamvaziripour, justinwu}@byu.edu, {zappala, seamons}@cs.byu.edu.

2471-2566/2019/04-ART13 $15.00
https://doi.org/10.1145/3313761

day use by novice users. This interest is likely spurred by concern over government surveillance of email, particularly when third-party services such as Gmail and Hotmail store email in plaintext on their servers. The Electronic Frontier Foundation has heavily promoted secure communication and has released a security scorecard of secure messaging systems that includes several email tools.[1]

Although Signal, WhatsApp, and other secure instant messaging platforms are becoming popular, it is unclear whether efforts to encourage users to likewise switch to secure email will succeed, given that usable, secure email is still an unsolved problem more than 15 years after it was first formally studied (Whitten and Tygar 1999). Moreover, widespread use of secure email depends in part on whether it could be adopted in a grassroots fashion, where both parties of an email conversation are novice users. All prior laboratory usability studies of secure email bring one novice participant at a time into the lab and have him or her communicate with a study coordinator using a secure email system. Although this helps researchers understand how well a novice can start using secure email when paired with an expert, it does not shed light on whether a pair of novices can start using the system independently.

In this work, we describe a novel paired-participant methodology for the study of secure email, in which pairs of novice participants were brought into the lab and asked to exchange secure email between themselves. We asked participants to bring a friend with them to ensure that the participants already knew each other and might behave more naturally. Participants then used one or more secure email systems without any specific training or instructions on how to use the system other than what the system itself provided. The main differences between this type of study and a traditional single-participant study are that the participants played different roles (initiating contact vs. being introduced to secure email) and that they interacted with another novice participant and not a study coordinator.

In this article, we describe two different studies using this methodology. In our first study, 28 pairs of participants tested three different secure email systems: Private WebMail (Pwm), Tutanota, and Virtru. Each of these systems represents a different integration strategy of secure email with existing email systems. Pwm integrates secure email with users' existing Gmail accounts, allowing them to compose and receive secure email with a familiar interface. In contrast, Tutanota is a secure email depot that requires users to log into Tutanota's website to interact with their secure messages. Virtru is a hybrid of these two approaches, allowing users who install the Virtru plugin to use secure email that is integrated with Gmail, but also allowing non-Virtru users to receive encrypted email through a depot-based system on Virtru's website.

In our second study, 10 pairs of participants tested Mailvelope, a modern Pretty Good Privacy (PGP) tool that was designed with usability in mind. Mailvelope is a browser extension that integrates with users' webmail systems and is the only such tool that appears on the EFF's secure messaging scorecard.

Our results and participant comments lead to the following contributions:

(1) **Using pairs of novice participants for an email usability study has several benefits.** Having participants play different roles allowed us to gather data about different types of first-use cases (i.e., sending a secure email first vs. receiving a secure email first). In addition, participants exhibited more natural behaviors and indicated that they felt "more at ease," that they and their friend were "on the same page" or at the same level of technical expertise, and that they did not feel discomfort from being "under the microscope."
(2) **Participants prefer integrated solutions over depot-based solutions.** Although to some it may be intuitive that users would prefer to continue using their existing email accounts, a number of depot-based systems have appeared recently (e.g., Tutanota,

[1] https://web.archive.org/web/20150909023035/https://www.eff.org/secure-messaging-scorecard.

ProtonMail, StartMail). Most of our participants strongly dislike using separate websites such as secure email depots to read their email.

(3) **Tutorials improve the usability of secure email.** When asked what they liked about Pwm and Virtru, participants often reported that they appreciated the tutorials presented alongside these systems. The efficacy of these tutorials is shown by the fact that while using Pwm and Virtru, participants were able to quickly complete the study task, whereas while using Tutanota—which lacks a tutorial—participants took on average 72% longer to complete the study tasks, often making mistakes as they did so.

(4) **Participants want the ability to use secure email but are unsure about when they would use it.** Three-fourths of the participants in our study indicated that they wanted to be able to encrypt their email, but only one-fourth indicated that they would want to do so frequently. Furthermore, when asked to describe how they would use encrypted email in practice, most participants were unsure, giving only vague references to how secure email might be useful. This demonstrates a need for future research to establish whether the true problem facing the adoption of secure email is usability or some other reason, such as day-to-day users having no need to send sensitive data via email and not understanding the risks of sending sensitive data via email.

(5) **PGP with manual key management still appears to be unusable for the masses.** In our second study, 9 of the 10 participant pairs were unable to complete the study and had not even made significant progress in the hour allotted for the study. The only pair that completed the study took slightly longer than the allotted hour and reported that they were only successful at using the tool because one of them had learned about public key cryptography in a college course. Moreover, all participants found key establishment and sharing difficult.

(6) **PGP-based secure email systems can be improved.** Based on our observations during both studies and on participants' feedback, we have identified several suggestions that would increase the usability of PGP-based secure email. First, integrated tutorials would be helpful in assisting first-time users in knowing what they should be doing at any given point in time. Second, software should automatically generate key pairs for users, prompting them only for a password to encrypt their generated private key. Third, the PGP block itself could be enhanced with human-readable text, to help users who receive an encrypted email and do not have the proper key to decrypt it to move forward. Fourth, software should assist users by automating the exchange of public keys and generating invitation emails that instruct recipients on how to install the PGP software.

Artifacts. The study materials and data described in this article are available for download at https://isrl.byu.edu/data/tops2018/. For convenience, we also include the study materials in the Appendix.

2 BACKGROUND

In this section, we provide background on the area of secure email. First, we describe the current state of email security. We then describe approaches for securing email with end-to-end encryption. Finally, we discuss usability studies of secure email.

2.1 Email Security

When email was first designed in 1971,[2] no meaningful attention was paid to security. As such, it was trivial for an attacker to steal email during transit or to send messages with falsified sender

[2] https://openmap.bbn.com/~tomlinso/ray/firstemailframe.html.

information. More recently, there have been attempts to patch security into email. For example, TLS is now used to protect email during transmission, and SPF, DKIM, and DMARC are used to authenticate the sending domain of an email. However, the deployment of these technologies is limited, and they are often misconfigured.

In an analysis of email delivery security (i.e., TLS, DKIM, DMARC, SPF), Durumeric et al. (2015) found that a majority of email is still vulnerable to attack. They showed that only 35% of SMTP servers are configured to use TLS, and even when TLS is enabled, it is often vulnerable to a downgrade attack. Similarly, they demonstrated that the adoption of DKIM and DMARC were so low that they provide no practical benefits. These results were further confirmed in concurrent work by both Foster et al. (2015) and Holz et al. (2016).

As such, email is still an easy target for attackers. For example, Durumeric et al. (2015) found that in seven countries, over 20% of inbound Gmail messages were being stolen. Additionally, the inability to authenticate the sender of an email increases the likelihood of email phishing, a multibillion-dollar problem.[3] Perhaps most troubling, even if TLS, DKIM, and DMARC were to be widely adopted and configured correctly, these would do nothing to protect email at rest, where email can be compromised as the result of a breach, a malicious insider, or a subpoena.

2.2 End-to-End Email Encryption

The problems afflicting secure email could be solved through the use of end-to-end encryption. A well-known approach for providing end-to-end encryption is PGP (Garfinkel 1995). It was developed in 1991 by Phil Zimmerman and allows users to encrypt and sign their email messages using public key cryptography. In PGP, keys are generally validated using the web of trust—for instance, users verify and sign their associates' public keys, and the users can check if they trust a key by seeing if they or one of their associates has signed that key. Public keys can be shared in a number of ways, such as sending the key directly to other users (i.e., manual key management), posting the key to a personal website, or uploading the key to a key directory.

Another approach using end-to-end encryption is Secure/Multipurpose Internet Mail Extensions (S/MIME) (Ramsdell and Turner 2010). Similar to PGP, S/MIME uses public key cryptography to encrypt and sign email. Unlike PGP's web of trust, S/MIME certificates are authenticated using the same certificate authority (CA) system used for authenticating websites with TLS. Users still need to share public keys just as they do in PGP. More recent versions of PGP also support certificates validated using the CA system, but in our experience this feature is not widely used.

Practical attacks against both PGP and S/MIME were recently demonstrated that exploit obsolete cryptographic primitives to coerce many email clients to utilize a backchannel to exfiltrate the plaintext of encrypted emails (Poddebniak et al. 2018). This revealed a major weakness in both the standards and the deployment of the two most mature technologies for end-to-end encrypted email.

A third approach is Identity-Based Encryption (IBE) (Shamir 1984). IBE is a public key system where a user's public key is deterministically derived from his or her email address. Private keys are generated by a trusted third-party server, which authenticates the identity of the user before providing the user with his or her private key. With IBE, senders can encrypt a message for any recipient without any prior setup or coordination with the recipient. Unlike PGP and S/MIME, users do not need to install secure email software, generate a key pair, and share their public key before they can receive their first encrypted message.

In addition to public key cryptography, it is also possible to use symmetric key cryptography for end-to-end encryption of email. One approach is for a user to select and share a password that

[3] https://krebsonsecurity.com/2016/04/fbi-2-3-billion-lost-to-ceo-email-scams/.

will be used to encrypt his or her email. Another approach is to have a trusted key escrow server that can generate and distribute symmetric keys for users.

In this article, we analyze the usability of systems that represent a variety of end-to-end encryption approaches: PGP with manual key management, IBE, passwords, and a custom key escrow scheme.

2.3 Usability Studies of Secure Email

There were two early usability studies of PGP and S/MIME. Whitten and Tygar (1999) conducted the first formal user study of a secure email system (i.e., PGP 5), which uncovered serious usability issues with key management and users' understanding of the underlying public key cryptography. They found that a majority of users were unable to successfully send encrypted email in the context of a hypothetical political campaign scenario. The results of the study served as a wake-up call to the security community and helped shape modern usable security research.

Next, Garfinkel and Miller (2005) created a secure email system using S/MIME and used this system to replicate Whitten and Tygar's earlier study. Their work demonstrated that automatic key management significantly increases usability compared to manual key management. Still, they observed that their tool "was a little too transparent" in how well it integrated with Outlook Express, and sometimes users failed to read the instructions accompanying the visual indicators. Since these two studies, there have been no usability studies of commercial email systems until our study.

Several researchers have built secure email prototypes or mockups that have proven usable in laboratory studies. Ruoti et al. (2013) developed Pwm, a secure email tool that uses IBE to encrypt users' messages. Across three usability studies of Pwm, they demonstrated that all participants could use Pwm to send and receive encrypted email. However, they also found that some users made mistakes and did not fully understand how the system worked due to the transparency of its automatic encryption. They later revised Pwm to address these issues and demonstrated that their modified system receives the highest usability ratings of any secure email system tested in the literature (Ruoti et al. 2016b).

Atwater et al. (2015) also evaluated the usability of PGP using a mockup of a secure email tool that automatically generates key pairs for users, shares the generated public key with a key server, and retrieves the recipient's public key as needed. Their results showed that with these modifications, users could successfully use PGP to send and receive secure email. Unfortunately, their mockup did not correctly simulate PGP's key management, making it difficult to determine whether the usability gains they observed generalize to a more correct implementation.

In a similar vein, Bai et al. (2016) explored user attitudes toward different models for obtaining a recipient's public key in PGP. In their study, they built two PGP-based secure email systems: one that used manual key exchange, in which users must directly exchange their public keys with each other (or use a web of trust), and one that used a key directory. Users were then provided with instructions on how to use each tool and given several tasks to complete. Afterward, participants shared their opinions regarding the key exchange models. The results of this study showed that overall, individuals preferred the key directory, although they were not averse to manual key exchange. Relatedly, Lerner et al. (2017) built Confidante, a secure email tool that leverages Keybase, a public key directory, for key management. A user study of Confidante with lawyers and journalists demonstrated that these users could quickly and correctly use the system.

In all of the studies described previously, participants interacted with an expert user of secure email (i.e., the study coordinator). In contrast, the methodology in this work is unique in that it tests whether *two* novice users can collaboratively begin using secure email. As discussed in this article, this approach has advantages over using a study coordinator to simulate a user.

Finally, some researchers have forgone the use of any research prototypes or mockups and instead directly surveyed users. Gaw et al. (2006) interviewed users at a political activist organization that use secure email and noted that adoption was driven by the organization deciding encryption was necessary due to secrecy concerns. They found that having IT staff set up the secure email software was necessary to enable the successful adoption of secure email. Still, even with this support, there were users who did not intend to use the software regularly, due to usability concerns and social factors. Renaud et al. (2014) and Abu-Salma et al. (2017) both surveyed users and found that most users did not have an accurate mental model regarding how email functioned and how encrypted email might help the user. Renaud et al. suggest that this lack of understanding made it difficult for users to understand the value of secure email, potentially explaining why users have not adopted it. Abu-Salma et al. note that the lack of interoperability between the various secure email clients is a barrier to adoption; this is in contrast to the instant messaging space where the smaller number of client applications make it easier to deploy end-to-end encryption that is largely invisible to the end users.

2.4 Relationship to Prior Publication

This article extends an earlier study that we published at CHI 2016 (Ruoti et al. 2016a). That work describes a within-subjects user study wherein pairs of participants used three secure email tools (Pwm, Tutanota, and Virtru) to communicate sensitive information to each other. When analyzing the results of this within-subjects study, we detected a flaw in our assignment of treatments (i.e., the order systems were used), which had a statistically significant impact on results. A bug in the Qualtrics survey software led to an uneven distribution of treatments, so Virtru was the first system that was used in two-thirds of the study sessions, potentially biasing the results. To address this limitation, we replicated the within-subjects study, correcting the flaw in our treatment assignment. The choice of which secure email tools to use, the task design and instructions, and the study questionnaire were identical between the two studies. Our replication study is described in Sections 4 and 5.

Concurrent with the original within-subjects study, we also conducted a separate study of Mailvelope using the same methodology. We did not include it in the study with the three other tools because pilot studies showed that the Mailvelope task would take too long to include with three other tools in a within-subjects test. The Mailvelope study is described in Sections 6 and 7. The materials for both studies can be found in Appendixes A and B.

3 SECURE EMAIL

Two and a half decades after the invention of PGP, secure email still remains sparsely used. Although some businesses require the use of secure email by their employees, secure email has not been adopted by the population at large. Although it is possible that secure email will eventually diffuse from the workplace, it may be that if secure email is to flourish, it will do so because of grassroots adoption—ordinary people discovering secure email on their own and easily beginning to use it with their acquaintances.

Previous secure email studies have not evaluated usability within the framework of a grassroots adoption paradigm. Instead, they evaluated interactions between a single novice and a study coordinator, who was an expert user. Both the study of Whitten and Tygar (1999) and the study of Garfinkel and Miller (2005) used a simulated political campaign, where the study participant was the only individual in the campaign who did not already know how to use PGP. Similarly, studies by Sheng et al. (2006), Ruoti et al. (2013), Song (2014), Atwater et al. (2015), and Bai et al. (2016) involved participants sending email to study coordinators, none of whom were instructed to simulate a novice user.

Even if the study coordinators had attempted to simulate a novice user, there are difficulties with this approach. First, study coordinators are unlikely to make mistakes while using the encryption software, which is atypical of a true novice. Even if study coordinators make use of scripted mistakes, there is a strong risk that these mistakes might be seen as artificial by participants, thereby breaking immersion for the participant. Second, in many tasks, there is a high level of possible variability in participant actions, making it difficult to script for all possible situations, and unscripted responses from coordinators are likely to be biased by their experience with the system. Third, participants are likely to attribute any problems they encounter to their own mistakes, and not to the coordinator, whereas when interacting with a friend, participants are just as likely to attribute the mistake to their friend as to themselves.[4]

To avoid these difficulties, our study uses two novice participants. This study tests whether two novice participants, who know each other beforehand, can successfully use secure email without any aid. Success in this scenario would indicate that grassroots adoption is feasible. Our observations, as discussed later in this article, show that this approach produces more natural behavior than when participants email a study coordinator. Moreover, this approach allows us to examine how users perform when they are introduced to secure email in different ways (i.e., installing and then sending an email vs. receiving an email and then installing).

To select which systems to test, we surveyed existing secure email systems, including those listed on the EFF's scorecard, and filtered them according to two criteria. First, we focused on webmail solutions, as previous work has shown that this approach is preferred by users (Ruoti et al. 2013; Atwater et al. 2015). Second, we required the systems to use automatic key management, as research has shown that users are highly amenable to this approach (Garfinkel and Miller 2005; Ruoti et al. 2013). Of the systems that matched these criteria, we found that they could be grouped into three types of secure email systems: integrated, depot-based, and a hybrid of integrated and depot-based systems. For each of these groups, we personally evaluated the systems in the group and selected the system that we felt had the best usability for inclusion in our study. In addition to the preceding systems, we also included a secure email system based on PGP, because this approach is viewed as highly secure by the research community and a recent system claimed to significantly improve its usability.

The remainder of this section describes the types of secure email that were tested, as well as the representative system for each.

3.1 Integrated Secure Email (Pwm)

Integrated secure email refers to secure email systems that integrate with users' existing email systems. In this model, users do not need to create new accounts and are able to encrypt messages within the email interfaces they are already accustomed to Ruoti et al. (2013).

Pwm is the representative system for this type of secure email. Pwm was developed as part of our research (Ruoti et al. 2013, 2016b) and received the highest System Usability Scale (SUS) scores (Brooke 1996) of all the secure email system tested in the literature (Ruoti et al. 2016b; Atwater et al. 2015) using that scale. In addition, because Pwm has previously been the subject of several formal user studies, it provides a good baseline for comparing the results of the other systems tested in this study.

Pwm is a browser extension that tightly integrates with Gmail's web interface to provide secure email. Users are never exposed to any cryptographic operations, including the verification of the user's identity, which are completed without user interaction. Pwm provides a secure composition

[4]In some ongoing work, we have attempted to simulate a novice user and encountered these difficulties in practice (Ruoti et al. 2016b).

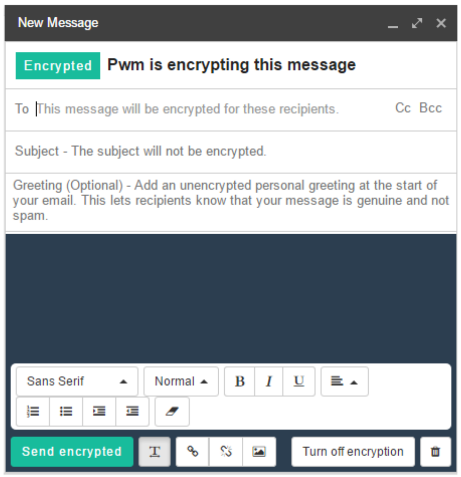

Fig. 1. Pwm: Secure composition interface.

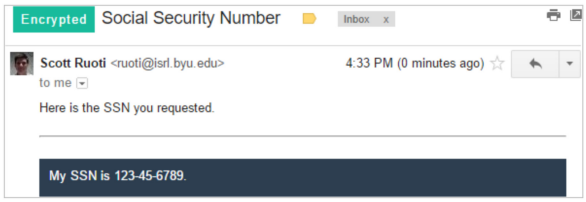

Fig. 2. Pwm: Secure read interface.

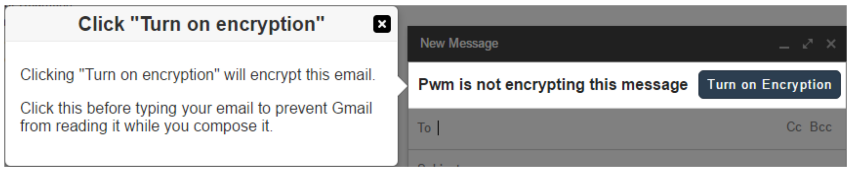

Fig. 3. Pwm: Integrated tutorial.

interface that shields plaintext from Gmail (Figure 1) and modifies the color scheme of Gmail for encrypted emails to help users identify which messages have been encrypted (Figure 2). Pwm also includes inline tutorials that instruct users on how to operate Pwm (Figure 3).

Pwm's threat model is focused on protecting email from individuals who do not have access to the sender's or recipient's email account. Although this does not protect email against attackers who compromise the user's email account, it does provide security during transmission and storage of the email. Pwm is susceptible to an attacker that compromises the extension software. Pwm is also susceptible to a malicious email service provider that impersonates the user or uses social engineering to obtain sensitive data.

3.2 Depot-Based Secure Email (Tutanota)

Depot-based secure email refers to secure email systems that use a separate website from users' existing email systems. In this model, users have a separate account with the depot where they can send and receive secure emails. When a user receives a new message in their depot account, many depot-based systems will send an email to the user's standard email address, informing the user that he or she has a new email to check in the depot system. Often, these systems do not allow users to send secure email to individuals not already using the depot, or they send these recipients a link to a secure message that can be read at the depot. Depot-based systems are commonly deployed by companies and organizations for secure communication.

There are many depot-based systems to choose from. We chose Tutanota[5] because it was the most usable of the depot systems we tested, was available for free, was available to new users,

[5]https://tutanota.com/.

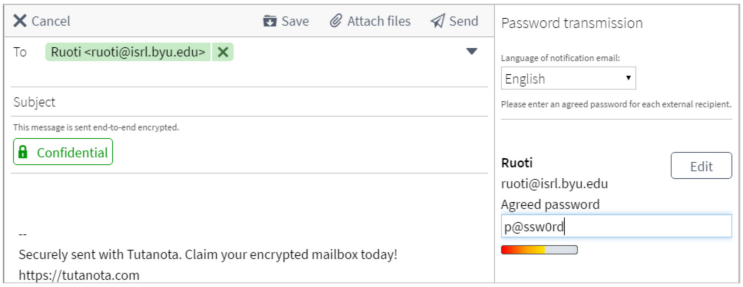

Fig. 4. Tutanota: Choosing a password for a non-Tutanota recipient.

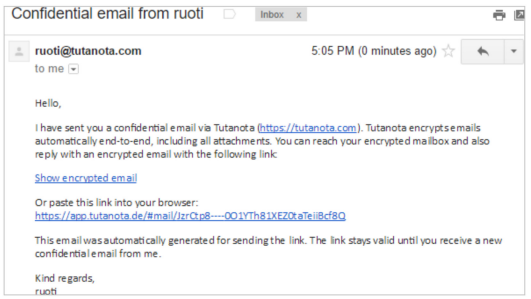

Fig. 5. Tutanota: A non-Tutanota recipient receiving a notification.

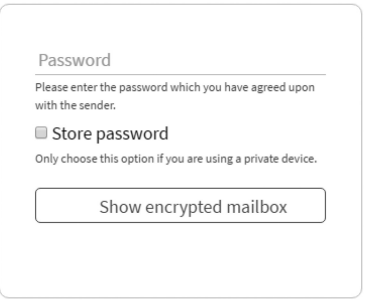

Fig. 6. Tutanota: Entering a password to read an encrypted email.

and was receiving positive publicity on Twitter. Other popular systems charged an annual fee (e.g., Hushmail and StartMail) or were currently not offering email addresses to new users (e.g., ProtonMail).

Tutanota assigns users an email address ending in `@tutanota.com`. Users can send and receive email from this address as they normally would. During account creation, Tutanota generates a public/private key pair for the user. These keys are stored on Tutanota's servers, with the private key being encrypted with the user's Tutanota account password. When Tutanota users send messages to other Tutanota users, the messages are automatically encrypted and signed with the appropriate keys. Because all of the key management is hidden from the user, Tutanota's email interface looks similar to any other webmail system. All cryptographic operations are carried out by the user's browser, to try to shield private key material from Tutanota's servers.

When a Tutanota user sends a message to a non-Tutanota user, the sender has the option of encrypting it with a password (Figure 4). When the non-Tutanota user receives the encrypted email (Figure 5), the user is redirected to Tutanota's website, where he or she can enter the password and decrypt the message (Figure 6). Tutanota's interface also allows the non-Tutanota user to respond to the message and will encrypt the response using the same password.

The threat model for Tutanota is similar to Pwm, except instead of having normal and secure emails stored in the same email accounts, they are stored in separate accounts. This means that if a user's normal email account is compromised, his or her sensitive messages are still secure. Users are also susceptible to a malicious email service provider that provides software to access the user's data, or to having their secure email account password guessed/stolen.

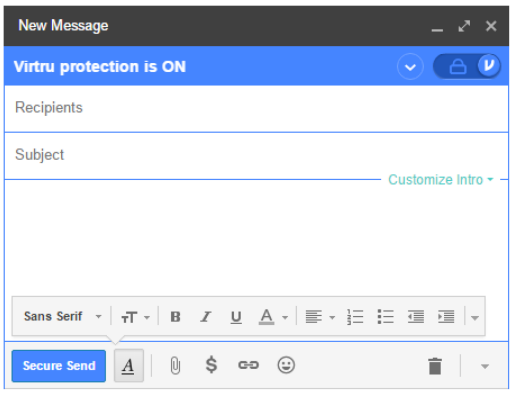

Fig. 7. Virtru: Secure composition interface.

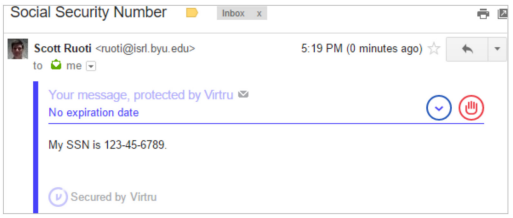

Fig. 8. Virtru: Secure read interface.

3.3 Hybrid Secure Email (Virtru)

Virtru[6] is a hybrid of integrated and depot-based secure email. Once Virtru's browser plugin is installed, it functions much the same as Pwm, including automatic key management and integration with the webmail provider (Figures 7 and 8). If a Virtru user sends an email to a non-Virtru user, the sender still does so through the webmail provider, but the recipient will receive an email informing the recipient that he or she needs to log into Virtru's website to view the message. At this point, Virtru is similar to Tutanota in its management of new users, except instead of providing a password, non-Virtru users are asked to prove that they own their email address. Virtru also includes tutorials, similar to Pwm.

The threat model for Virtru is identical to Pwm.

3.4 PGP (Mailvelope)

We chose Mailvelope,[7] a modern PGP-based tool, as our representative system. Mailvelope was our preferred choice for several reasons. First, it is the only PGP-based secure email tool promoted by the EFF's secure messaging scorecard that uses webmail. Second, Mailvelope is highly rated on the Chrome Web Store, with 242 users collectively giving it 4.6 of 5 stars. Third, Mailvelope claims to be "easy-to-use" and focused on helping novice users begin sending encrypted email.[8] Finally, in our evaluation of other PGP-based secure email tools, we found Mailvelope to be at least as usable as the alternatives (i.e., GPG Tools, Enigmail, Google's End-to-End Encryption).

Like Pwm, Mailvelope integrates with the user's webmail provider. Upon installation, users need to generate a PGP key pair and select a password to encrypt their private key. To encrypt a message, the user takes the following steps: (1) click on the button that opens Mailvelope's compose interface in a new window; (2) compose the message (Figure 9), click encrypt, and select the intended recipients (Figure 10); and (3) click the transfer button, which then sends the user's PGP-encrypted message back to the webmail provider's compose interface, where the user can then send the encrypted message (Figure 11). Upon receipt of an encrypted email, the user takes the following steps: (1) click the lock icon that is displayed over the encrypted text (Figure 12), and (2) enter the password for his or her private key (Figure 13).

[6] https://www.virtru.com/.
[7] https://www.mailvelope.com/.
[8] For the first claim, see https://www.mailvelope.com/faq, and for the focus of the project, see https://github.com/mailvelope/mailvelope/issues/14#issuecomment-11419791.

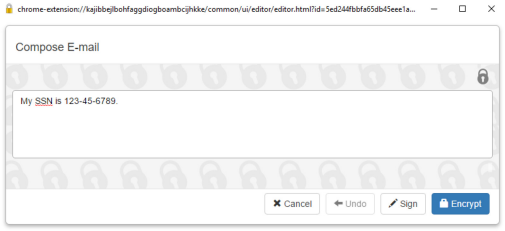

Fig. 9. Mailvelope: Secure composition interface.

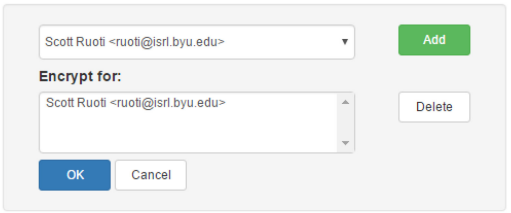

Fig. 10. Mailvelope: Public key selection interface.

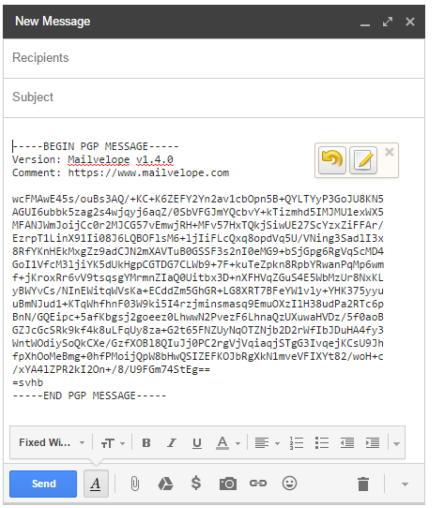

Fig. 11. Mailvelope: Encrypted message in the webmail provider's compose interface.

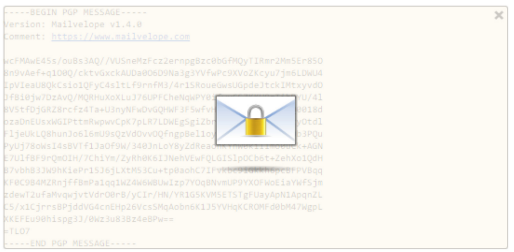

Fig. 12. Mailvelope's encrypted message in the webmail provider's read interface.

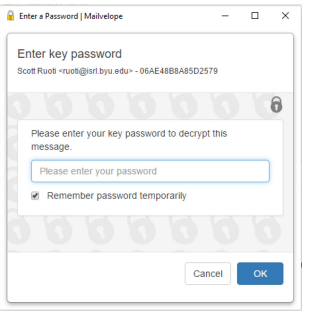

Fig. 13. Mailvelope's private key password entry interface.

Mailvelope has a stricter threat model than the other systems. To compromise a message encrypted with PGP, the attacker must accomplish three things: (1) steal the user's email, (2) steal the user's private key, and (3) steal the password for the user's private key. An attacker who gains access to the user's email account could attempt to convince the user's contacts to encrypt messages with the attacker's public key instead of the user's true public key. Still, this does not compromise the security of messages previously encrypted with the correct public key.

4 WITHIN-SUBJECTS STUDY: METHODOLOGY

In this section, we discuss the methodology of the replicated within-subjects study that corrects the ordering error from our original study (Ruoti et al. 2016a).

4.1 Study Setup

The study ran for just over three weeks—beginning Monday, February 13, 2017, and ending Wednesday, March 8, 2017. In total, 28 pairs of participants (56 total participants) completed the

study. Participants took between 30 and 60 minutes to complete the study, and each participant was compensated $15 USD for participating. Participants were required to be accompanied by a friend, who served as their counterpart for the study. For standardization and requirements of the systems tested in the study, both participants were required to have Gmail accounts.

When participants arrived, they were read a brief introduction detailing the study and their rights as participants. Participants were informed that they would be in separate rooms during the study and would use email to communicate with each other. The study coordinators ensured that the participants knew each other's email addresses. Participants were also informed that a study coordinator would be with them at all times and could answer any questions they might have.

Using a coin flip, one participant was randomly assigned as Participant A (referred to as "Johnny" throughout the article) and the other as Participant B (referred to as "Jane" throughout the article). The participants were then led to the appropriate room to begin the study; each room had identical equipment. For the remainder of the study, all instructions were provided in written form. Participants completed the task on a virtual machine, which was restored to a common snapshot after each study task, ensuring that the computer started in the same state for all participants and that no participant information was accidentally stored.

During the study, participants were asked to complete a multistage task three times, once for each of the secure email systems being tested: Pwm, Tutanota, and Virtru. An enumeration of the six possible task orderings was created and shuffled. Each participant was sequentially assigned to one of the task orderings. To facilitate timely completion of each study session, participants were given 10 minutes to complete each task. Study coordinators were instructed to allow extra time to participants as necessary, but if participants were unable to complete the task within a reasonable time limit, the study coordinators marked the task as incomplete and moved on to the next system.

4.2 Demographics

We recruited Gmail users at Brigham Young University for our study. Participants were almost evenly split between male and female: female (27; 48%), male (29; 52%). Participants skewed young: 18 to 24 years old (45; 80%), 25 to 34 years old (11; 20%). Most participants rated their level of computer expertise at "Intermediate": Beginner (10; 18%), Intermediate (44; 78%), Advanced (2; 4%).

We distributed posters across campus to avoid biasing our results to any particular major. Almost all of the participants were university students: students (54; 96%), non-students (2; 4%). Student participants were enrolled in a variety of majors, including both technical and non-technical majors.

4.3 Scenario Design

During the study, participants were asked to role-play a scenario about completing taxes. Each participant was shown the following text:

- **Johnny.** Your friend graduated in accounting and you have asked their help in preparing your taxes. They told you that they needed you to email them your last year's tax PIN and your social security number. Since this information is sensitive, you want to protect (encrypt) this information when you send it over email.
- **Jane.** You graduated in accounting and have agreed to help a friend prepare their taxes. You have asked them to email you their last year's tax PIN and their social security number.

Participants were provided with the information they would send (e.g., SSN and PIN) but were told to treat this information as they would their own sensitive information.

4.4 Task Design

Based on the scenario, participants were asked to complete a three-stage task:

(1) Johnny would encrypt and send his SSN and last year's tax PIN to Jane.
(2) Jane would reply to this sensitive information with a confirmation code and this year's tax PIN. This information would also be encrypted.
(3) Johnny would reply and let Jane know he had received the confirmation code and this year's tax PIN.

The instructions guiding the participants through the three stages are as follows:

— **Johnny.** In this task, you'll be using {Pwm, Virtru, or Mailvelope}. The system can be found at the following website: {Appropriate website}. Please encrypt and send the following information to your friend using {Pwm, Virtru, or Mailvelope}: SSN: {Generated SSN}. PIN: {Generated PIN}.

  Once you have received the confirmation code and PIN from your friend, send an email to your friend letting them know you have received this information. After you have sent this confirmation email, let the study coordinator know you have finished this task.

— **Jane: Sheet 1.** Please wait for your friend's email with their last year's tax PIN and SSN. Once you have written down your friend's SSN and PIN, let the study coordinator know that you are ready to reply to your friend with their confirmation code and PIN.

— **Jane: Sheet 2.** You have completed your friend's taxes and need to send them the confirmation code and this year's tax PIN from their tax submission. Since your friend used {Pwm, Virtru, or Tutanota} to send sensitive information to you, please also use {Pwm, Virtru, or Tutanota} to send them the confirmation code and PIN. Confirmation Code: {Generated code}. PIN: {Generated PIN}.

  Once you have sent the confirmation code and PIN to your friend, wait for them to reply to you and confirm they received the information. Once you have received this confirmation, let the study coordinator know you have finished this task.

The instructions for Johnny and Sheet 1 of the instructions for Jane were given at the start of the task. Sheet 2 for Jane was given once Jane had received and decrypted the sensitive information sent by Johnny in Stage 1. Participants completed this task once for each of the three systems being tested. Each time, the instructions only included information relevant to the system being tested.

While participants waited for email from each other, they were told that they could browse the Internet, use their phones, or engage in other similar activities. This was done to provide a more natural setting for the participants, and to avoid frustration if participants had to wait for an extended period of time while their friend figured out an encrypted email system.

Study coordinators were allowed to answer questions related to the study but were not allowed to provide instructions on how to use any of the systems being tested. If participants became stuck and asked for help, they were told that they should take whatever steps they normally would to solve a similar problem. Additionally, when asked for help, if the study coordinator believed communication between the two parties could help, he could remind participants that they were free to communicate with their friend and that only the sensitive information was required to be transmitted over secure email.

Study coordinators observed the participants to ensure that the task was completed successfully—for instance, Jane received the encrypted tax information from Johnny. If participants failed to install the needed software or transmitted the required information in plaintext, the study coordinator would remind the user of his or her intended task. Once the coordinator observed that

the task was complete or the 10 minutes allocated to completing the task ran out, the coordinator would close the virtual machine and return the participant to the study questionnaire.

4.5 Study Questionnaire

We administered our study using the Qualtrics web-based survey software. Before beginning the survey, participants answered a set of demographic questions. Participants then completed the study task for each of the three secure email systems.

Immediately upon completing the study task for a given secure email system, participants were asked several questions related to their experience with that system. First, participants completed the 10 questions from the SUS (Brooke 1996, 2013) that are listed in our survey reproduced in Appendix B. Studies have shown that SUS is a good indicator of perceived usability (Tullis and Stetson 2004) and is consistent across populations (Ruoti et al. 2015). It has been used in the past to rate secure email systems (Ruoti et al. 2013; Atwater et al. 2015). After providing a SUS score, participants were asked to describe what they liked about each system, what they would change, and why they would change it.

After completing the task and questions for all three secure email systems, participants were asked to select which of the encrypted email systems they had used was their favorite, and to describe why they liked this system. Participants were next asked to rate the following statements using a five-point Likert scale (Strongly Disagree to Strongly Agree): "I want to be able to encrypt my email" and "I would encrypt email frequently."

4.6 Post-Study Interview

After completing the survey, participants were interviewed by their respective study coordinator. The coordinator asked participants about their general impressions of the study and the secure email systems they had used. Furthermore, the coordinators were instructed to note when the participants struggled or had other interesting events occur, and during the post-study interview the coordinators reviewed and further explored these events with the participants.

After the participants completed their individual post-study interviews, they were brought together for a final post-study interview. First, participants were once again asked which system was their favorite and why. This question was intended to observe how participants' preferences might change when they could discuss their favorite system with each other. Second, participants were asked to describe their ideal secure email system. Although participants are not system designers, our experience has shown that this question often elicits preferences that otherwise remain unspoken. Finally, participants were asked to share their opinions related to doing a study with a friend. This question was designed to learn possible benefits and limitations of conducting such a two-person study.

4.7 Limitations

Our study has limitations common to all existing secure email studies. First, our study population was predominantly young students who were experienced computer users. The results are not generalizable, and future research should include more diversity. Second, our study was a short-term study, and future research should explore these issues in a longer-term longitudinal study. Third, our study is a lab study and has limitations common to all studies run in a trusted environment (Milgram and Van den Haag 1978; Sotirakopoulos et al. 2010).

Our study only examines the case where a single user sends email to one other user. Although this is the most likely scenario for secure email among the masses, future work could also explore alternative scenarios (e.g., sending to multiple users, mailing lists).

Table 1. SUS Scores

	Participant	Count	Mean SUS Score	Standard Deviation	Confidence Interval ($\alpha = 0.05$)	Range
Pwm	Johnny	28	80.5	12.1	±4.5	76.1–85.0
	Jane	28	76.4	12.3	±4.6	71.9–81.0
	Both	56	**78.5**	12.1	±3.2	75.3–81.7
Virtru	Johnny	28	73.4	15.5	±5.7	67.7–79.1
	Jane	27	68.8	13.8	±5.2	63.6–74.0
	Both	55	**71.1**	14.6	±3.9	67.3–75.0
Tutanota	Johnny	28	50.2	17.2	±6.4	43.8–56.5
	Jane	24	49.3	17.1	±6.8	42.4–56.1
	Both	52	**49.8**	16.8	±4.6	45.2–54.3

SUS was designed as a measure of perceived usability, not an absolute measure of usability or security. For instance, it does not measure a user's understanding of risk or their perception of a system's overall security. As such, SUS should not be interpreted as a final judgment regarding the usable security of a system but rather as just one aspect of the overall usability and security assessment—for instance, a system could receive a high SUS score and still have significant security drawbacks. Still, when taken in conjunction with other measures of usability (e.g., task completion time) and security (e.g., number of mistakes), perceived usability measures—such as SUS—can be a helpful tool used to compare the perceived usability of competing systems. For example, perceived usability measures can help identify systems that users would be extremely unlikely to adopt in real life.

5 WITHIN-SUBJECTS STUDY: RESULTS

In this section, we report the quantitative results from our replicated within-subjects study. First, we report on perceived usability scores for each system. Next, we give the time taken to complete the task for each system as well as the number of mistakes encountered while using each system. Finally, we report which system participants indicated was their favorite. When multiple statistical tests were run on the same data, we adjusted p values using a Bonferroni correction.

5.1 System Usability Scale

We evaluated each system using the SUS to measure perceived usability. A breakdown of the SUS score for each system and type of participant (i.e., Participant A—Johnny, Participant B—Jane, or both) is given in Table 1 and Figure 14. The mean value for both participants is used as the system's final SUS score (Brooke 1996) and is displayed in bold in the table. Five participants' SUS ratings were discarded (four for Tutanota, one for Virtru) because they never used the system. This was because the Johnny participant paired with him or her was unable to send them a message within the time limit. Other scores from tasks that timed out before participants could complete them are included, because the participants were able to use the systems even if they did not finish the task.

To give greater context to the meaning of each system's SUS score, we leveraged the work of several researchers. Bangor et al. (2009) analyzed 2,324 SUS surveys and derived a set of acceptability ranges that describe whether a system with a given score is acceptable to users in terms of usability. Bangor et al. also associated specific SUS scores with adjective descriptions of a system's

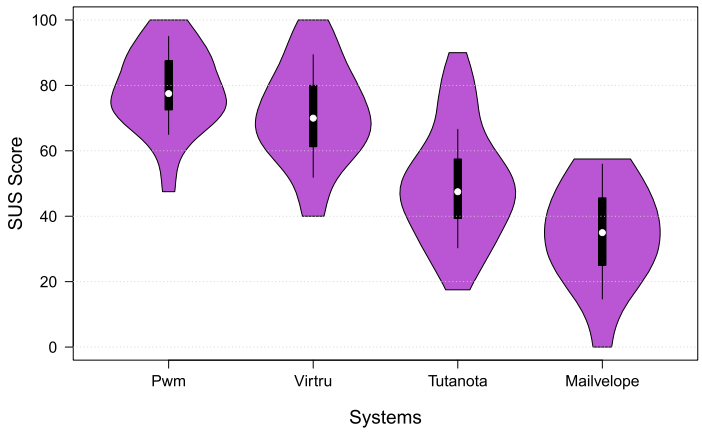

Fig. 14. Violin plot of SUS scores from the within-subjects study and the Mailvelope study.

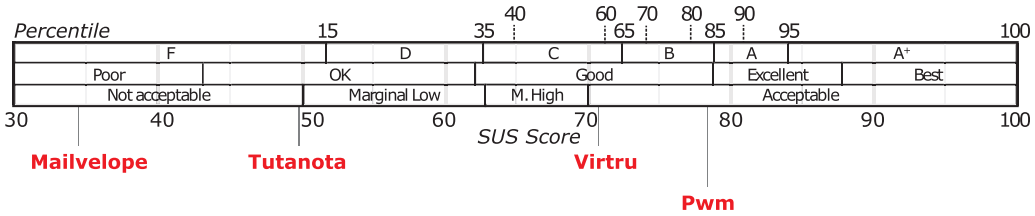

Fig. 15. Adjective-based ratings and percentiles to help contextualize SUS scores.

perceived usability. Using this data, we generated ranges for these adjective ratings such that a score is correlated with the adjective it is closest to in terms of standard deviations. Sauro (2011) also analyzed SUS scores from Bangor et al. (2008), Tullis and Stetson (2004), and their own data. They calculated the percentile values for SUS scores and assigned letter grades based on percentile ranges. This data from Bangor et al. and Sauro et al. is graphically represented in Figure 15.

Pwm's SUS score of 78.5 falls just below the 85th percentile, placing it at the top of the "Good" range and "B" grade, whereas Virtru's score of 71.1 is under the 60th percentile in the "C" grade. The scores for Pwm are roughly consistent with those seen in prior work (Ruoti et al. 2013, 2016b). Tutanota's score of 49.2 falls below the 15th percentile into the "F" grade. This system is rated as having "OK" usability. A one-way repeated measures ANOVA showed an omnibus effect ($F(2, 100) = 76.02, p < .001$), with Tukey's HSD test showing that the difference between all three systems was statistically significant (Pwm and Virtru—$p < 0.05$, Pwm and Tutanota—$p < 0.01$, Virtru and Tutanota—$p < 0.01$).

5.2 Time

We measured the time it took each participant to finish the task. Times were measured from the recorded video of each study session and reflect the total time it took for both participants to complete the task. Timing began when Johnny was given the first instruction sheet and ended when Johnny sent the reply email to Jane that he had received the confirmation code and PIN. Study coordinators allowed participants to slightly exceed the 10-minute time limit if it looked like participants were likely to finish with a few additional minutes. If after 10 minutes had elapsed participants had not made any significant progress, the coordinators marked the task as incomplete and had the participants move to the survey portion of the task.

Table 2. Time Taken to Complete Task (min:sec)

	Count	Incomplete	Completion Percentage	Mean Task Time	Standard Deviation	Confidence Interval ($\alpha = 0.05$)	Range
Pwm	28	0	100%	9:16	2:44	±1:01	8:15–10:17
Virtru	26	2	93%	8:37	2:56	±1:07	7:29–09:44
Tutanota	17	11	61%	10:49	2:23	±1:08	9:41–11:57

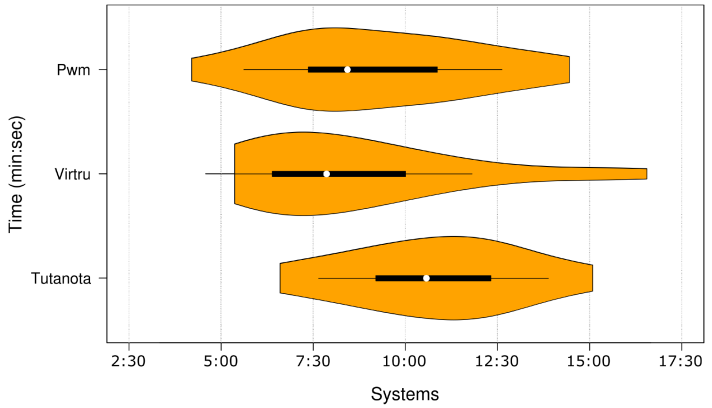

Fig. 16. Violin plot of completion times.

The task times are reported in Table 2 along with the number of incomplete tasks in which participants were unable to complete the task in the given time limit. Incomplete tasks are not factored into the mean time or related statistics. Figure 16 shows a violin plot of task completion times. A one-way repeated measures ANOVA showed an omnibus effect ($F(2, 30) = 5.67, p < .01$), with Tukey's HSD test showing that the difference between Tutanota's task completion time and the other two systems was statistically significant (Pwm and Virtru—$p > 0.05$, Pwm and Tutanota—$p < 0.05$, Virtru and Tutanota—$p < 0.01$).

Tutanota had the highest rate of task failure (11 pairs; 39%). In most cases, this failure was caused by one of two issues:

(1) Time ran out before Johnny finished registering with Tutanota and sending an encrypted email to Jane.
(2) Johnny sent the encrypted email to Jane, but they were unable to successfully share the password before running out of time.

5.3 Mistakes

We defined mistakes as any situation in which sensitive information was sent in plaintext or was sent encrypted along with the key to decrypt the sensitive information (i.e., the Tutanota password used to encrypt the email was sent as plaintext in email). Failure to complete the task within the time limit was not counted as a mistake. Overall, five participant pairs (five pairs; 18%) failed to enable encryption before sending sensitive information: one pair for Pwm (one pair; 4%), three pairs for Virtru (three pairs; 10%), and one pair for Tutanota (one pair; 4%). Additionally, two

Table 3. Participants' Favorite Systems

	Johnny		Jane		Total	
Pwm	16	(57%)	19	(68%)	35	(63%)
Virtru	10	(36%)	6	(21%)	16	(28%)
Tutanota	2	(7%)	3	(11%)	5	(9%)

participant pairs (two pairs; 7%) using Tutanota successful encrypted the sensitive information, but then sent the password for that email over standard, non-encrypted email.

5.4 Favorite System

At the end of the study, participants were asked which of the three systems was their favorite. Their responses are summarized in Table 3. Pwm was most frequently rated as the favorite system, with Virtru also rated highly. Tutanota was rarely selected as the favorite system. These results roughly correlate with the SUS score of each system. Although they were given the option, no participants indicated that they disliked all of the systems.

Interestingly, we do see a difference in the choice of favorite system based on what role the participant played. Although Pwm is rated as the favorite system more often than Virtru by both Johnny and Jane, the difference in favorite systems is more extreme for Jane. Based on participant responses, this disparity is due to the fact that unlike Johnny, Jane had to leave Gmail to interact with Virtru messages, a process that was frequently described negatively.

Similarly, Tutanota was more highly rated by Jane than by Johnny. Participant responses reveal that this is likely due to the fact that Jane did not have to go through the Tutanota account setup (which required a long, complex password) and selection of a password to encrypt the email (which caused nearly all participants to struggle).

5.5 Comparison to the Original Study

As discussed in Section 4, the within-subjects study described in this article is a replication of an original study in which system orderings were not evenly distributed (Ruoti et al. 2016a). There were two differences between the results of the two studies:

(1) **Pwm received a higher SUS score in the replicated study.** In the original study, Pwm had received a mean SUS score of 72.7, whereas in the replicated study, it received a mean SUS score of 78.5. This latter score is more in line with previous evaluations of Pwm (Ruoti et al. 2016b). This difference was expected as an analysis of results from the original study suggested that system ordering had an affect on SUS scores.

(2) **Increased task failure rate in the replicated study.** In the original study, nearly all participants completed all tasks. In the replicated study, one-third of participants failed to complete the Tutanota task, and two participant pairs failed to complete the Virtru task. This difference can be attributed to study coordinators in the replication study being more stringent about ending the tasks when participants had not completed the task within 10 minutes. In contrast, study coordinators in the original study never ended a task and in some cases let participants take 20 minutes to finish using a system.

    The difference in task failure rates also affected the mean task completion times for each of the systems. In the original study, Pwm and Virtru had similar mean task completion times—with Pwm slightly faster—and Tutanota had a mean task completion time nearly twice as long as Pwm or Virtru. In the replicated study, all three systems had similar task completion times. This difference is easily explained by the lack of

long-running task times for Virtru and Tutanota caused by study coordinators ending tasks instead of letting them take longer to complete.

(3) **Different mistake rates.** In the original study, no participants made any mistakes using Pwm, compared to a single mistake in this study. Similarly, in the original study, no participants failed to encrypt their messages using Tutanota, compared to a single such mistake in this study. Positively, in the original study, two-thirds of participants using Tutanota sent their password over an unsafe channel, whereas in the replicated study, only two participant pairs (two pairs; 7%) made this mistake. Virtru had a single mistake in both the original and replicated studies.

6 MAILVELOPE STUDY: METHODOLOGY

Using the same methodology as the within-subjects study, we conducted an IRB-approved user study wherein pairs of participants used Mailvelope to transmit sensitive information to each other.

6.1 Study Setup

The study ran for 2 weeks, beginning Tuesday, September 8, 2015, and ending Friday, September 18, 2015. In total, 10 pairs of participants (20 total participants) completed the study. Unlike the within-subjects study, participants only used a single system—Mailvelope. Participants were allocated 60 minutes to complete the study, with about 35 to 40 minutes spent using Mailvelope.

6.2 Demographics

We recruited Gmail users at Brigham Young University for our study. Participants were two-thirds male: male (13; 65%), female (7; 35%). Participants skewed young: 18 to 24 years old (18; 90%), 25 to 34 years old (2; 10%).

We distributed posters broadly across campus to avoid biasing our results to any particular major. All participants were university students,[9] with the majority being undergraduate students: undergraduate students (17; 85%), graduate students (3; 15%). Participants were enrolled in a variety of majors, including both technical and non-technical fields of study. No major was represented by more than four participants, with most having only one or two participants.

6.3 Limitations

Limitations from the first study also apply here. Additionally, this second study only included 20 participants, all of whom were students. Although this was enough to show difficulties associated with Mailvelope, it is not indicative of all possible outcomes. It would be especially interesting to rerun this study using different populations (e.g., technical professionals, computer scientists, security professionals).

7 MAILVELOPE STUDY: RESULTS

Overall, participants were unable to use Mailvelope to send encrypted email, with only 10 in 10 participant pairs completing the assigned task within the provided hour. This is in stark contrast to the results of the within-subjects study in which most participants completed the assigned tasks in under 10 minutes, and all participants would likely have completed the tasks if given 45 minutes per system. As such, it is clear that Mailvelope is not suitable for helping novices send secure email between themselves.

In the remainder of this section, we detail Mailvelope's failure rate, report on its SUS scores, and list mistakes made by participants.

[9]We did not require this.

Table 4. Mailvelope SUS Scores

	Count	Mean	Standard Deviation	Confidence Interval ($\alpha = 0.05$)	Range
Johnny	10	30.5	16.6	±10.3	20.2–40.8
Jane	6	41.3	10.9	±8.7	32.6–50.0
Both	16	34.5	15.3	±7.5	27.0–42.0

7.1 Failures

The study lasted an hour, with roughly 45 minutes allocated to complete the assigned task. If, after installing Mailvelope, Johnny made absolutely no progress for 30 minutes, study coordinators were instructed to end the task and continue to the post-study interview. If instead participants were making some progress, study coordinators would allow them to continue until there were 10 minutes left in the hour, reserving the last 10 minutes for the post-study interview.

Of the 10 participant pairs, 9 were unable to successfully complete the task. In 2 of those 9 pairs, Johnny never figured out how to use Mailvelope to send any message.[10] In another 2 pairs, Jane was completely mystified by the encrypted PGP message and was unaware that she needed to install Mailvelope to read it. Only 1 of the 9 pairs traded public keys, although this pair was still confused about what to do after sharing their public keys.

The one pair that completed the task required more than the full 45 minutes to do so. At their request, we allowed 10 extra minutes to complete the task. We did so because this was the only pair that appeared close to finishing the task, and we were interested in observing a successful trial. Interestingly, in this pair, Jane indicated that they previously learned about public key cryptography in a college class and attributed their success to this prior knowledge. Based on our observation of this pair, we agree that Jane's knowledge of public key cryptography was instrumental to the pair's success at completing the assigned task.

7.2 System Usability Scale

We evaluated Mailvelope using the SUS to obtain a measure of its perceived usability. Although Johnny always completed the SUS evaluation for Mailvelope, in four instances Jane never progressed far enough in the assigned task to install the system, and so we did not have her complete the SUS questions. A breakdown of the SUS score for each type of participant is given in Table 4. Differences between the SUS scores for all four systems tested can be seen in Figure 14.

Mailvelope's SUS score of 34.5 is rated as having "Poor" usability. It falls in the 4th percentile, is given a letter grade of "F," and is labeled as "Not acceptable." The differences between Mailvelope and the three systems tested in the first study are all statistically significant (two-tailed student t-test, equal variance, Bonferroni correction applied—$p < 0.001$).

7.3 Mistakes

All participant pairs made mistakes. The most common mistake was encrypting a message with the sender's public key. This occurred for seven of the participant pairs, including the participant pair that was eventually successful. Three of the participant pairs generated a key pair with their

[10] In these two instances, the study was stopped after 30 minutes because no progress had been made by Johnny. The other eight pairs were given the full 45 minutes to complete the task.

friend's information and then tried to use that public key to encrypt their message. One participant modified the encrypted message after encryption (while still in Mailvelope's compose window), adding their sensitive information to the area before the PGP block. Finally, one participant eventually exported his private key and sent it along with his key ring password to his friend so that his friend could decrypt the message he had received. In this case, even though the participants had transmitted the required information, they were informed that they needed to try some more and accomplish the task without sending the private key.

8 QUALITATIVE RESULTS AND DISCUSSION

In this section, we discuss themes that we noticed across both studies, especially the qualitative feedback provided by participants on the study survey and in the post-study interview. Participants in the first study were assigned a unique identifier R[1–28][A, B], and participants in the second study were also assigned a unique identifier M[1–10][A, B]. The final letter refers to which role the participant played during the study, and participants with the same number were paired with each other (e.g., R1A and R1B were Johnny and Jane, respectively, in the same study session).

8.1 Paired-Participant Methodology

During the studies, we noticed several clear benefits of our paired-participant methodology.

First, by having participants play different roles, we were able to gather data about participant experiences both when they are introduced to secure email and when someone else is introducing them. For example, in Tutanota, messages need to have a password to be encrypted. Johnny's experiences revealed the difficulty in discovering that a password is required and that it needs to be communicated to the recipient—Jane. Similarly, Jane's experience showed the aversion participants felt to leaving their current email system to view a sensitive message. Although these same experiences might have been elicited by running two different studies, it was convenient to obtain them in a single study. In particular, when errors were encountered, we found it helpful to compare what both sides (Johnny and Jane) reported about the error. In addition, showing that a participant can successfully use a new secure email system when inducted by another novice user is a stronger indication of its ease of adoption, as compared to only showing that a new user can be inducted by an expert.

Second, our study design led to more natural behaviors by participants. In past studies, we observed that participants expected study coordinators to immediately respond to emails. Even after being informed that a response would take several minutes and that they could do other things while waiting, participants would constantly refresh their inbox to see if a message had arrived, and if a response took longer than 15 to 30 seconds to arrive, participants would often complain. In contrast, participants in our studies were content to wait to receive their email and did not appear agitated when their friends took a long time to respond. Instead of constantly refreshing their inbox, participants exhibited a variety of actions—responding to non-study related emails, browsing the Web, checking social media, doing homework, or using their phones—which is likely more representative of how they use email in practice.

As another example, in several instances when participant pairs would encounter a problem using one of the systems, one of the participants would blame the problem on his or her partner and refuse to take any action other than telling the partner to go and "figure it out." This happened regardless of which partner was at fault. In contrast, when problems would occur in our single-participant studies, the participant would assume that it is his or her fault (even when it is not) and would immediately try to figure out how to remedy the problem.

In addition to the observations by study coordinators, participants also noted that they felt more natural interacting with a friend than with a study coordinator. For example, participant R18B said

that doing the study with a friend was "convenient, because we knew how to verify each other" and that verifying the identity of a stranger would be harder. R12B stated that sending encrypted email to a friend "made it feel more applicable" to how he or she would use secure email in the real world. Furthermore, participants benefited from being able to communicate out-of-band from the secure email tool, similar to how two novices would communicate in a real-life use case. Participant R8B stated,

> "[It was] convenient because he could text me. [We] knew each other well enough that that was a normal mode of communication."

Some participants indicated that because they were working with their friends, they felt more relaxed. For example, participant R20A said, "[I] definitely felt more comfortable. With a stranger it would be kind of awkward." Mistakes are common for novice users of any system, but working with a friend allowed our participants to not become discouraged when they made mistakes. R9B said, "I don't feel as stressed when I don't get it right." R3A and R3B both expressed a similar sentiment, adding that they could laugh at their mistakes instead of becoming distressed. Finally, we were pleased to note that requiring participants to bring a friend with them resulted in a much lower missed appointment rate than we have seen in the past.

It should be noted that not all participants thought the paired-participant methodology was beneficial. As such, 86% of participants had a positive opinion of the methodology; the rest were ambivalent, stating that doing the study with someone they knew did not make a difference. None of the participants had an explicitly negative opinion of the methodology.

Based on our observation of participants' behavior and the participants' qualitative feedback, we believe that there is significant value in conducting two-person studies. Still, future research should examine in greater depth the differences between one- and two-person studies. For example, an A/B study comparing these two methodologies could be conducted that compares differences in system metrics (e.g., SUS, task completion time), as well measures differences in users' agitation during the study (e.g., heart rate, eye tracking). Similarly, research could compare how participant experiences differ when both roles are filled by a novice as opposed to having one simulated by a coordinator.

Finally, we note that the two-person methodology described here does not remove the need for other study setups. For example, several key motivating examples for secure email (e.g., whistle-blowing) involve using secure email between users who are strangers. In these cases, the traditional single-participant methodology or a two-person methodology using strangers would be more appropriate.

8.2 Passwords

Tutanota supports password-based encryption, whereas Pwm and Virtru use email-based identification and authentication (Garfinkel 2003) to verify the user's identity to a key escrow server. The password used by Tutanota to encrypt an email made it clear that only the recipient who had the password would be able to read the message. This well-understood form of security made a large number of participants feel that Tutanota was the most secure system, even if usability issues prevented it from being their favorite system. R1B's response demonstrates this principle:

> "It felt more secure. The fact that you had to use a password made it seem more secure. So if I was really worried about my security—like if I was in politics or something and I was really worried about it—then I think I'd go with [Tutanota] because the password would make it feel more secure."

The positive opinion toward passwords was so strong that several participants stated that they wished Pwm and Virtru would also allow them to password-encrypt messages. For example, R9A said,

> "[Pwm and Virtru] need more security, for example link to Gmail but each time you send it you have the password. Because if Gmail is hacked, it's kinda pointless."

Still, not all participants were enamored with using a password to encrypt email, seeing it as an added memory burden or hassle. As stated respectively by participants R4B and R16A,

> "I have mixed feeling about the shared password mechanism. It is kind of nifty but also something of a hassle. Adds a layer of security that I might like though."

> "[Passwords were] one step more than was needed…I didn't like having to make my friend contact me to get a password."

Multiple participants (14%, n = 8) pointed out that requiring passwords raises the security question of how to transmit those passwords safely. Two participants suggested that they would prefer a system that allows them to choose whether to encrypt with a password or not on a message-by-message basis.

8.3 Security Indicators

Some participants were concerned that it was impossible to verify if any of the systems were truly encrypting their data. This likely stems from two facts: first, that participants are not security experts and lack the means to truly verify the security of a tool, and second, that the tools themselves—once working—do not show the user sufficient indication that the email he or she is receiving is actually encrypted. Results from Atwater et al. (2015) and Ruoti et al. (2016b) suggest that showing ciphertext does not address this issue, and the fact that participants are concerned indicates that this problem needs more research. For example, participants R6B and R16B respectively stated,

> "I just need something that would give me reassurance that it couldn't just be accessed by anyone."

> "(Survey Question: What would you change about Pwm?) Maybe a way of showing that the encryption worked and for sure no one else could see it?"

Many participants mentioned security indicators such as lock icons that differentiated encrypted and unencrypted email. Some participants felt that the existing indicators were sufficient to assure them that their emails were encrypted, whereas others pointed to the appearance of encrypted email as something they would change about the systems they used. For example, participants R4B and R15A respectively said,

> "[Regarding Pwm and Virtru:] If the email just looked more different, I would feel more secure."

> "[Regarding Virtru:] I would make it more obvious when sending an email through gmail if you were sending an encrypted email or an normal email."

8.4 Privacy of Encrypted Email

More than one-half (59%, n = 33) of the study participants believed that only the sender and intended recipient could read an encrypted email. Of those whose believed that someone other than the sender and receiver could read the email as well (23%, n = 13), three-fourths included the email

provider or encryption system developers among those who could decrypt their messages. This demonstrates a lack of understanding of the role the encryption key plays in cryptography or at least how the key is handled. This lack of understanding continues to be a problem for secure email systems (Ruoti et al. 2013).

Another interesting belief held by multiple participants was in the existence of "hackers" who are able to break passwords or decrypt email through the application of skill alone. For example, participants R9A and R23B respectively stated,

> "(Interviewer: Do you think there's anyone besides the recipient that can read your encrypted email?) Probably a lot of other people who have means to hack into the account."

> "(Interviewer: Who do you think can read the encrypted emails you sent?) Maybe if there was some crazy hacker person like on TV. They always seem to be able to access anything. Other than that, I don't think so."

Some participants expressed distrust in encrypted email because they believed it was susceptible to such adversaries. Participant R21B said that he would never send sensitive information over the Internet because "there are people know how to decrypt things, even. So even at that point it's not safe." The same participant also considered out-of-band transmission of secrets (e.g., the shared password required by Tutanota) via SMS or voice call to be unsafe and vulnerable to compromise.

8.5 Integrated Versus Depot Systems

Participants overwhelmingly preferred secure email to be integrated into their existing email systems and not require a second account (i.e., a depot). This preference was evident by the low SUS scores of Tutanota and the fact that only five participants rated it as their favorite system. These results are in line with previous work also showing that users prefer integrated systems (Ruoti et al. 2013; Atwater et al. 2015).

Additionally, participant comments made it clear that they were not interested in using depot-based secure email. For example, participants R12B, R15A, and R11A respectively stated,

> "Another email [account] is the last thing I would want."

> "Nobody wants to create a separate email account to send private information."

> "I had to create a Tutanota account before I could send an encrypted email which is time-consuming. I would want something which enables me to use my already existing email accounts to do the same."

However, a minority of participants felt that Tutanota was more secure than other systems, precisely because it required the creation of an account separate from Gmail.

Users' preference for integrated secure email is also shown in participant interactions with Virtru. When users have Virtru installed on their machine, they can read and compose messages within Gmail. In contrast, when non-Virtru users receive a Virtru-encrypted message, the message does not prompt recipients to download and install the Virtru plugin but instead takes them to an external webpage with message-depot functionality, bypassing the Gmail integration that participants were so fond of. This was disliked by several participants, with participant R6B including "[switching] back and forth between Virtru and Gmail to receive messages" among the reasons she disliked Virtru.

Finally, we note that a depot-based system that did not allow communication with outside parties would remove the possibilities of mistakenly sending information in the clear. Still, it is unclear whether such a system is still an email system and not yet another specialized chat/communication

application. In addition, our results show that users may be unwilling to adopt such a system, illustrating a tension between usability and security that exists between the integrated and depot-based system designs.

8.6 Tutorials

Tutorials were a significant factor in participants' experiences, providing further evidence for the claim that tutorials are key to Pwm's high usability (Ruoti et al. 2016b). Pwm was rated by participants as having the best tutorials, with more than one-fourth of participants (29%, n = 16) bringing up tutorials when asked what they liked about Pwm. Participants largely liked the style of the tutorials as well as their content. For example, participant R1A expressed, "I liked that it gave very simple step by step instructions."

Virtru also has tutorials, but praise for these tutorials was not as common as it was with Pwm, with Jane participants criticizing the tutorials more than Johnny participants. This result can likely be attributed to the fact that the Virtru plugin walks new users through a tutorial upon installation, but someone who receives a Virtru-encrypted message without the plugin is simply presented with a blue button labeled "Unlock message" without additional instruction beyond what the sender of the email has personally and manually added. This is in contrast to Pwm, which prefaces incoming encrypted email with instructions on what encrypted email is and how the recipient should go about decrypting the message.

Tutanota had no tutorials, and this clearly led to confusion. Many participants failed to notice that they needed to set a password to encrypt their email, and just as many did not realize that they needed to communicate this password to the other participant. Additionally, some participants did not understand that they could not just use Tutanota to communicate the password. Many of these problems could have been alleviated by a simple tutorial, and several participants mentioned that Tutanota would be improved by the addition of a tutorial.

8.7 Reasons to Use Encrypted Email

The majority of participants (70%) agreed with the survey statement, "I want to be able to encrypt my email," although only a much smaller fraction (29%) agreed with the statement that they would "encrypt email frequently." Some participants described scenarios when they could see themselves using encrypted email, such as sending sensitive information or keeping business transactions private. Participant R4B said that even with encrypted email, They would still only send sensitive information to a close family member. A few participants stated that they could not ever envision themselves using secure email. Their reasons fell into one of two categories. Either they felt they did not have information that was sensitive enough to warrant encryption, or they did not trust the security of encrypted email. These sentiments respectively were expressed by R18B and R10A:

> "If we ever wanted to exchange information like that, we would just do it in person."

> "I don't really send a lot of secret information…I don't think I would use it a lot."

8.8 Guidelines for Usable PGP

Mailvelope clearly failed to help the majority of participants encrypt their email—only 1 in 10 pairs succeeded. All participants expressed frustration with Mailvelope, with the most comical expression of this frustration coming from M3A: "Imagine the stupidest software you would ever use, and that was what I was doing." The difficulty also led several participants to indicate that in the real world they would have given up trying to use Mailvelope long before they did during the study. For example, M3A also said, "After five minutes, I would have just given up and called."

The accumulated evidence over many years indicates that it is not feasible to expect novice users to understand public key cryptography to use PGP software, nor to directly manage keys. It is still an open question whether PGP-based secure email can be made sufficiently usable for the masses, without requiring specialized knowledge. Generally, evidence from prior studies indicates that automating and simplifying user interactions with public key cryptography is likely to have success. Thus, we believe that the following guidelines will help PGP-based secure email tools be significantly more usable for novice users.

8.8.1 Integrated Tutorials. When using Mailvelope, participants were constantly flipping between Mailvelope's website and Gmail, looking for instructions on what to do next. At no stage was it intuitive how they should proceed based on Mailvelope's UI. Nearly all participants indicated that they wished Mailvelope had provided instructions that were integrated with the Mailvelope software and would walk them through, step-by-step, in setting up Mailvelope and sending their first encrypted email. As seen for Pwm and Virtru, tutorials likely could greatly assist first-time users in acclimating to PGP.

Important steps that could be addressed by tutorials are (1) explanation of any user interactions required when generating key pairs, (2) inviting their friends to set up a secure email tool, (3) sharing public keys, (4) sending their first encrypted email, (5) decrypting their first encrypted email, and (6) backing up key material.

8.8.2 Automatic Key Generation. Johnny participants struggled to generate their own PGP key pair, with much of the confusion tied to their lack of understanding regarding public and private keys. Often participants were unsure whose information they should input into the key generation dialog—their own or that of the intended participants. This was only compounded by the fact that Mailvelope showed users their own public key in the list of "recipients" the message was encrypted for. An easy way to address this problem would be to automatically generate a user's key pair during installation, retrieving necessary information from the webmail provider (i.e., name and email address), and only prompting the user for the private key's encryption password. The software should also not allow operations with keys that are nonsensical, such as encrypting messages with your own private key, or including a private key in an outgoing email.

8.8.3 Better Text to Accompany PGP Block. During Johnny's attempts to send Jane an encrypted email, Johnny often encrypted a message for himself and then sent that encrypted message to Jane. Upon seeing the PGP ciphertext block, Jane was unclear what she was supposed to do with it. One participant noted that she thought it was an image that had gotten garbled during email transmission.

Although Johnny had obviously made a mistake, it also represented an opportunity for Jane to recognize that Johnny was trying to use secure email. To make better use of this opportunity, the PGP ciphertext block could be modified to include read-only plaintext instructions detailing the nature of the encrypted email, how to obtain a PGP-based secure email tool, and how to start sending encrypted email (Ruoti et al. 2013). Although this would not allow Jane to read the email from Johnny, it would allow her to better collaborate with Johnny in discovering how to use secure email. An indication that this approach could be successful is given by participant M9B, the only Jane participant who finished the study task. In referring to the PGP ciphertext block, she said, "It was like a puzzle, I only got a link to Mailvelope. I then had to go there and explore."

8.8.4 Automatic Key Discovery and Email Invites. Similarly, Johnny participants were confused about what Jane needed to do to receive an encrypted email. Much of this confusion was centered around how to share keys. To address this, we recommend the use of an automated key exchange through a trusted directory, as is done with Pwm; through Keybase, as is done with

Confidante (Lerner et al. 2017); an auditable ledger, as with CONIKS (Melara et al. 2015); or directly through email. Although there are security tradeoffs with each of these approaches, they would address a significant hurdle for adoption of PGP-based secure email.

We also recommend that PGP clients detect when recipients do not have a public key and help the sender take the appropriate steps to resolve this problem. For example, a PGP client could generate an email for the recipient stating that the sender wants to communicate with him or her using PGP. This generated email could also include instructions on how to set up the PGP client. This technique was first explored by Atwater et al. (2015), and our experience leads us to strongly recommend it.

9 CONCLUSION

In this work, we conducted the first two-person study of secure email where two novice users are brought into the lab together and asked to exchange secure email between themselves. Our study analyzed Pwm, Tutanota, Virtru, and Mailvelope. Using a two-person study enabled us to observe participants under different first-use experiences. In addition, participants exhibited more natural behaviors, seemed less agitated, and indicated that they felt less like they were "under the microscope."

Our results indicate several observations about secure email systems. First, we found that participants largely reject depot-based secure email systems. Second, participant success in using a system without mistakes is heavily influenced by the presence of well-designed tutorials. Finally, although participants are interested in using secure email, few express a desire to use it regularly, and most are unsure of when or how they would use it in practice.

Our results also demonstrate that, after two and a half decades, PGP-based secure email using manual key management is still unsuitable for novice users. By comparing results from Mailvelope with the results from Pwm and Virtru that showed fewer mistakes and a higher perceived usability scores, we created several guidelines to help make PGP-based tools generally more suitable for novice users, principally through the use of automating and simplifying user interactions with public key cryptography.

Avenues for future work include the following:

(1) Quantifying the differences between two-person and single-person studies with respect to participant comfort and task completion.
(2) Building a usable, fully functional PGP system for manual and automatic key exchange using the guidelines presented in this article.
(3) An A/B evaluation of key management schemes that is not confounded by differing implementation details in competing industrial products.[11]
(4) Longitudinal studies of secure email to determine if the usability progress made in this and related work holds up over extended usage.
(5) The mistake rate for Pwm is low (zero mistakes in the original study, one in the replicated study). Still, to estimate the true mistake rate, it will be necessary to conduct studies with much larger populations (hundreds of pairs of users). Establishing the true mistake rate will be important, as even a relatively low mistake rate (e.g., 2%) might be completely unacceptable for some use cases and populations.
(6) Studies that broaden the populations used to test secure email.
(7) Since the SUS was designed for general usability and not usable security, future research should examine whether we can create better quantitative metrics for assessing the

[11] In follow-on research, we have begun exploring items (2) and (3) (Ruoti et al. 2018).

usability of secure systems. Additionally, alternative perceived usability measures (e.g., QUIS, SUMI, PSSUQ, SEQ) could be evaluated to determine if they are more suitable than the SUS for usable security studies.

APPENDIX

A RECRUITING POSTER

Fig. 17. Recruitment poster used to find participants.

B STUDY SURVEYS

B.1 Study Survey: Johnny

Survey Number *1; 2; 3; 4; 5; 6*

Demographics
What is your gender? *Male; Female; I prefer not to answer*

What is your age? *18–24 years old; 25–34 years old; 35–44 years old; 45–54 years old; 55+ years or older; I prefer not to answer*

What is the highest degree or level of education you have completed? *Some high school, no high school diploma; High school graduate, diploma or the equivalent (for example: GED); Some college or university credit, no degree; College or university degree; Post-Secondary Education; I prefer not to answer*

What is your occupation or major? *Free response*

How would you rate your level of computer expertise? *Beginner; Intermediate; Advanced*

Scenario

In this study, you will be role playing the following scenario:

Your friend graduated in accounting and you have asked their help in preparing your taxes. They told you that they needed you to email them your last year's tax PIN and your social security number. Since this information is sensitive, you want to protect (encrypt) this information when you send it over email.

You will be asked to send this information using three different secure email systems. In each task, you'll be told which system to use and assigned a new PIN and SSN. After correctly sending the information, your friend will reply to you with a confirmation code that can be used to continue with the study.

Task Instructions, Email Task, and Task Evaluation are seen three times, once for each system: Pwm, Virtru, and Tutanota

Task Instructions (Repeated three times)

Tell the study coordinator that you are ready to begin this task.
System: **Pwm**

In this task, you'll be using **Pwm**. The system can be found at the following website: **https://pwm.byu.edu/**.

Please encrypt and send the following information to your friend using **Pwm**:
SSN: 264-94-8748
PIN: 6567

Email Task (Repeated three times)

Enter the confirmation code provided by your friend. *Free response*

Enter the PIN provided by your friend. *Free response*

Task Evaluation (Repeated three times)
You will now be asked several questions concerning your experience with **Pwm**.

Please answer the following question about **Pwm**. Try to give your immediate reaction to each statement without pausing to think for a long time. Mark the middle column if you don't have a response to a particular statement.

Strongly Disagree; Disagree; Neither Agree nor Disagree; Agree; Strongly Agree

(1) I think that I would like to use this system frequently.
(2) I found the system unnecessarily complex.
(3) I thought the system was easy to use.
(4) I think that I would need the support of a technical person to be able to use this system.
(5) I found the various functions in this system were well integrated.
(6) I thought there was too much inconsistency in this system.
(7) I would imagine that most people would learn to use this system very quickly.
(8) I found the system very cumbersome to use.
(9) I felt very confident using the system.
(10) I needed to learn a lot of things before I could get going with this system.

What did you like most about using **Pwm**? *Free response*

What would you change about **Pwm**? *Free response*

Please explain why. *Free response*

Final Evaluation

You have finished all the tasks for this study. Please answer the following questions about your experience.

Which system was your favorite? *Virtru; Pwm; Tutanota*

Please explain why. *Free response*

Please answer the following question. Try to give your immediate reaction to each statement without pausing to think for a long time. Mark the middle column if you don't have a response to a particular statement.

Strongly Disagree; Disagree; Neither Agree nor Disagree; Agree; Strongly Agree

(1) I want to be able to encrypt my email.
(2) I would encrypt email frequently.

B.2 Study Survey: Jane

Survey Number *1; 2; 3; 4; 5; 6*

Demographics

What is your gender? *Male; Female; I prefer not to answer*

What is your age? *18–24 years old; 25–34 years old; 35–44 years old; 45–54 years old; 55+ years or older; I prefer not to answer*

What is the highest degree or level of education you have completed? *Some high school, no high school diploma; High school graduate, diploma or the equivalent (for example: GED); Some college or university credit, no degree; College or university degree; Post-Secondary Education; I prefer not to answer*

What is your occupation or major? *Free response*

How would you rate your level of computer expertise? *Beginner; Intermediate; Advanced*

Scenario

In this study, you will be role playing the following scenario:

You graduated in accounting and have agreed to help a friend prepare their taxes. You have asked them to email you their last year's tax PIN and their social security number.

As part of the study, your friend will send you this information three different times. Each time, after receiving their PIN and SSN, you will be provided with a confirmation code and a PIN number to send to your friend so that both of you can continue with the study.

Task Instructions, Email Task, and Task Evaluation are seen three times, once for each system: Pwm, Virtru, and Tutanota

Task Instructions (Repeated three times)

Tell the study coordinator that you are ready to begin this task.
System: **Virtru**

Email Task (Repeated three times)

Please wait for your friend's email with their last year's tax PIN and SSN.
Enter your friend's SSN. Include dashes. *Free response*

Enter your friend's PIN. *Free response*

Task Evaluation (Repeated three times)
You will now be asked several questions concerning your experience with **Virtru**.

Please answer the following question about **Virtru**. Try to give your immediate reaction to each statement without pausing to think for a long time. Mark the middle column if you don't have a response to a particular statement.

Strongly Disagree; Disagree; Neither Agree nor Disagree; Agree; Strongly Agree

(1) I think that I would like to use this system frequently.
(2) I found the system unnecessarily complex.
(3) I thought the system was easy to use.
(4) I think that I would need the support of a technical person to be able to use this system.
(5) I found the various functions in this system were well integrated.
(6) I thought there was too much inconsistency in this system.
(7) I would imagine that most people would learn to use this system very quickly.
(8) I found the system very cumbersome to use.
(9) I felt very confident using the system.
(10) I needed to learn a lot of things before I could get going with this system.

What did you like most about using **Virtru**? *Free response*

What would you change about **Virtru**? *Free response*

Please explain why. *Free response*

Final Evaluation

You have finished all the tasks for this study. Please answer the following questions about your experience.

Which system was your favorite? *Virtru; Pwm; Tutanota*

Please explain why. *Free response*

Please answer the following question. Try to give your immediate reaction to each statement without pausing to think for a long time. Mark the middle column if you don't have a response to a particular statement.

Strongly Disagree; Disagree; Neither Agree nor Disagree; Agree; Strongly Agree

(1) I want to be able to encrypt my email.
(2) I would encrypt email frequently.

REFERENCES

Ruba Abu-Salma, M. Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. 2017. Obstacles to the adoption of secure communication tools. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'17)*. IEEE, Los Alamitos, CA, 137–153.

Erinn Atwater, Cecylia Bocovich, Urs Hengartner, Ed Lank, and Ian Goldberg. 2015. Leading Johnny to water: Designing for usability and trust. In *Proceedings of the 1th Symposium on Usable Privacy and Security (SOUPS'15)*. 69–88.

Wei Bai, Moses Namara, Yichen Qian, Patrick Gage Kelley, Michelle L. Mazurek, and Doowon Kim. 2016. An inconvenient trust: User attitudes toward security and usability tradeoffs for key-directory encryption systems. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS'16)*. 113–130.

Aaron Bangor, Philip Kortum, and James Miller. 2008. An empirical evaluation of the system usability scale. *International Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies* 4, 3 (2009), 114–123.

John Brooke. 1996. SUS—A quick and dirty usability scale. In *Usability Evaluation in Industry*. CRC Press, Boca Raton, FL.

John Brooke. 2013. SUS: A retrospective. *Journal of Usability Studies* 8, 2 (2013), 29–40.

Zakir Durumeric, David Adrian, Ariana Mirian, James Kasten, Elie Bursztein, Nicolas Lidzborski, Kurt Thomas, et al. 2015. Neither snow nor rain nor MITM…: An empirical analysis of email delivery security. In *Proceedings of the 15th ACM Internet Measurement Conference (IMC'15)*. ACM, New York, NY, 27–39.

Ian D. Foster, Jon Larson, Max Masich, Alex C. Snoeren, Stefan Savage, and Kirill Levchenko. 2015. Security by any other name: On the effectiveness of provider based email security. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS'15)*. ACM, New York, NY, 450–464.

Simson Garfinkel. 1995. *PGP: Pretty Good Privacy*. O'Reilly Media, Inc., Sebastopol, CA.

Simson L. Garfinkel. 2003. Email-based identification and authentication: An alternative to PKI? In *Proceedings of the 24th IEEE Symposium on Security and Privacy (S&P'03)*. IEEE, Los Alamitos, CA, 20–26.

Simson L. Garfinkel and Robert C. Miller. 2005. Johnny 2: A user test of key continuity management with S/MIME and Outlook Express. In *Proceedings of the 1st Symposium on Usable Privacy and Security (SOUPS'05)*. ACM, New York, NY, 13–24.

Shirley Gaw, Edward W. Felten, and Patricia Fernandez-Kelly. 2006. Secrecy, flagging, and paranoia: Adoption criteria in encrypted email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'06)*. ACM, New York, NY, 591–600.

Ralph Holz, Johanna Amann, Olivier Mehani, Matthias Wachs, and Mohamed Ali Kaafar. 2016. TLS in the wild: An Internet-wide analysis of TLS-based protocols for electronic communication. In *Proceedings of the 24th Network and Distributed System Security Symposium (NDSS'16)*.

Ada Lerner, Eric Zeng, and Franziska Roesner. 2017. Confidante: Usable encrypted email: A case study with lawyers and journalists. In *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P'17)*. IEEE, Los Alamitos, CA, 385–400.

Marcela S. Melara, Aaron Blankstein, Joseph Bonneau, Michael J. Freedman, and Edward W. Felten. 2015. CONIKS: A privacy-preserving consistent key service for secure end-to-end communication. In *Proceedings of the 24th USENIX Security Symposium (USENIX Security'15)*. 383–398.

Stanley Milgram and Ernest Van den Haag. 1978. *Obedience to Authority*. Ziff–Davis Publishing Company, New York, NY.

Damian Poddebniak, Christian Dresen, Jens Müller, Fabian Ising, Sebastian Schinzel, Simon Friedberger, Juraj Somorovsky, et al. 2018. Efail: Breaking S/MIME and OpenPGP email encryption using exfiltration channels. In *Proceedings of the 27th USENIX Security Symposium (USENIX Security'18)*. 549–566. https://www.usenix.org/conference/usenixsecurity18/presentation/poddebniak.

B. Ramsdell and S. Turner. 2010. Secure/Multipurpose Internet Mail Extensions (S/MIME) Version 3.2 Message Specification. *RFC 5751 (Proposed Standard)*. Retrieved March 19, 2019 from http://www.ietf.org/rfc/rfc5751.txt.

Karen Renaud, Melanie Volkamer, and Arne Renkema-Padmos. 2014. Why doesn't Jane protect her privacy? In *Proceedings of the 14th Privacy Enhancing Technologies Symposium (PETS'14)*. 244–262.

Scott Ruoti, Jeff Andersen, Scott Heidbrink, Mark O'Neill, Elham Vaziripour, Justin Wu, Daniel Zappala, et al. 2016a. "We're on the same page": A usability study of secure email using pairs of novice users. In *Proceedings of the 34th ACM Conference on Human Factors and Computing Systems (CHI'16)*. ACM, New York, NY, 4298–4308.

Scott Ruoti, Jeff Andersen, Travis Hendershot, Daniel Zappala, and Kent Seamons. 2016b. Private webmail 2.0: Simple and easy-to-use secure email. In *Proceedings of the 29th ACM User Interface Software and Technology Symposium (UIST'16)*. ACM, New York, NY.

Scott Ruoti, Jeff Andersen, Tyler Monson, Daniel Zappala, and Kent Seamons. 2018. A comparative usability study of key management in secure email. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS'18)*. 375–394. https://www.usenix.org/conference/soups2018/presentation/ruoti.

Scott Ruoti, Nathan Kim, Ben Burgon, Timothy Van Der Horst, and Kent Seamons. 2013. Confused Johnny: When automatic encryption leads to confusion and mistakes. In *Proceedings of the 9th Symposium on Usable Privacy and Security (SOUPS'13)*. ACM, New York, NY.

Scott Ruoti, Brent Roberts, and Kent Seamons. 2015. Authentication melee: A usability analysis of seven web authentication systems. In *Proceedings of the 24th International Conference on World Wide Web (WWW'15)*. ACM, New York, NY, 916–926.

Jeff Sauro. 2011. *A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices*. Measuring Usability LLC, Denver, CO.

Adi Shamir. 1984. Identity-based cryptosystems and signature schemes. In *Proceedings of the 14th International Cryptology Conference (Crypto'84)*. 47–53.

S. Sheng, L. Broderick, C. A. Koranda, and J. J. Hyland. 2006. Why Johnny still can't encrypt: Evaluating the usability of email encryption software. In *Proceedings of the Poster Session at the Symposium on Usable Privacy and Security*.

Yuanzheng Song. 2014. *Browser-Based Manual Encryption*. Master's Thesis. Brigham Young University, Provo, UT.

Andreas Sotirakopoulos, Kirstie Hawkey, and Konstantin Beznosov. 2010. "I did it because I trusted you": Challenges with the study environment biasing participant behaviours. In *Proceedings of the Usable Security Experiment Reports Workshop at the Symposium on Usable Privacy and Security*.

Thomas S. Tullis and Jacqueline N. Stetson. 2004. A comparison of questionnaires for assessing website usability. In *Proceedings of the Usability Professional Association Conference*. 1–12.

A. Whitten and J. D. Tygar. 1999. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In *Proceedings of the 8th USENIX Security Symposium (USENIX Security'99)*. 14–28.

Received March 2018; revised January 2019; accepted February 2019