

Two-Factor or not Two-Factor?

A Comparative Usability Study of Two-Factor Authentication

Emiliano De Cristofaro¹, Honglu Du¹, Julien Freudiger¹, and Greg Norcie^{2,*}

¹ PARC (a Xerox Company), firstname.lastname@parc.com (authors contributed equally)

² Indiana University, greg@norcie.com

Abstract

Decades of research and numerous incidents have demonstrated the weaknesses of text passwords and prompted the need for more secure alternatives. In recent years, two-factor authentication (2F) has emerged as the most used solution to strengthen passwords. By requiring users to provide more than one authentication factor – e.g., a code generated by a security token, along with the password – 2F aims to enhance resilience against guessing attacks and breaches of password databases. Alas, it also introduces non-negligible costs for service providers and requires users to carry out additional actions during the authentication process, nevertheless, little research has focused on its usability.

This paper presents a comparative usability study of two-factor authentication. First, we report on a preliminary interview-based study involving 9 participants, identifying the most popular 2F technologies as well as the contexts and motivations in which they are used. Then, we design and administer a survey to 219 Mechanical Turk users, aiming to explore the landscape of 2F technologies and measure the usability of three popular solutions: codes generated by security tokens, one-time PINs received via email or SMS, and dedicated smartphone apps (e.g., Google Authenticator). We record contexts and motivations, and study their impact on perceived usability. We also present an exploratory factor analysis that captures some key factors affecting usability of 2F and highlight interesting findings that call for further research in the field.

1 Introduction

Today’s digital society increasingly depend on the ubiquitous availability of digital information and services. This prompts the need for effective mechanisms to securely manage online identities and authenticate users. Yet, despite its limited security, the most common method of authentication still relies on usernames and passwords. Guessing and dictionary attacks are often possible due to careless password choice: for instance, [28] shows that 20% of passwords are covered by a list of only 5,000 common passwords. To in-

crease their security, users have often little choice other than memorizing many different hard-to-guess passwords [34]. This creates an inherent conflict between security and usability, as analyzed by prior work, e.g., in the context of password selection [11], management [18, 32], and composition [20, 29].

As a consequence, multi-factor authentication technologies emerged as a way to enhance authentication security by requiring the user to provide more than one authentication *factor*, as opposed to only a password. Such authentication factors can be of three kinds:

1. *knowledge* – something the user knows, e.g., a password;
2. *possession* – something the user has, e.g., a security token (aka hardware token);
3. *inherence* – something the user is, e.g., a biometric characteristic.

In this paper, we concentrate on the most common instantiation of multi-factor authentication, i.e., the one based on two factors, which we denote as 2F. Historically, 2F has been deployed mostly in enterprise, government, and financial sectors, where sensitivity of information and services has driven institutions to accept increased implementation/maintenance costs, and/or to impose additional actions on authenticating users. In fact, in 2005, the United States’ Federal Financial Institutions Examination Council officially recommended the use of multi-factor authentication [9], thus pressuring most institutions to adopt some forms of two-factor authentication for online banking. Similarly, government agencies and enterprises often require employees to use 2F for, e.g., VPN authentication or B2B transactions. More recently, however, an increasing number of (non-financial) service providers—e.g., Google, Facebook, Dropbox, Twitter, GitHub, Evernote—have begun to provide their users with the option of enabling 2F, arguably, motivated by the increasing number of password databases recently compromised (and related bad press).¹

Nonetheless, 2F has a few limitations. For instance, as pointed out by Schneier [25], 2F technologies (including recently proposed ones based on fingerprints [26]) are often

*Work done while the author was at PARC.

¹E.g., Rockyou, Dropbox, Twitter, LinkedIn, Playstation Network.

vulnerable to man-in-the-middle, forgery, or Trojan-based attacks, and are not really effective against phishing. Also, 2F systems introduce non-negligible costs for service providers. Furthermore, several forms of 2F require users to carry out additional actions in order to authenticate, e.g., entering a one-time code and/or carrying an additional device with them. A common assumption in the IT sector, partially supported by prior work [4, 5, 6, 14, 23, 27], is that 2F technologies have low(er) usability compared to authentication based only on passwords, and this likely hinders larger adoption. Consequently, numerous start-up companies (e.g., PassBan, Duo Security, Authy, Encap, just to cite a few) aim to innovate the 2F landscape and introduce more usable solutions in the market.

Nonetheless, very little work thus far has comprehensively analyzed and compared the usability of different 2F technologies (see related work in next section). In this paper, we begin to address this gap by presenting a qualitative and quantitative analysis of 2F technology usability. First, we report on a preliminary, interview-based study involving 9 participants: we identify popular 2F technologies as well as the contexts and motivations in which they are used. Then, we design and administer a survey to 219 Mechanical Turk users, aiming to measure the usability of a few second-factor solutions: one-time codes generated by security tokens, one-time pins received via SMS or email, and dedicated smartphone apps (such as, Google Authenticator). We also record contexts and motivations, and study their impact on perceived usability of different 2F technologies.

Our comprehensive analysis of the landscape of 2F technologies yields some interesting findings. We show how users' perception of 2F usability is often correlated with their individual characteristics (such as, age, gender, background), rather than with the actual technology or the context in which it is used. We also present an exploratory factor analysis, which demonstrates that three metrics – ease-of-use, required cognitive efforts, and trustworthiness – are enough to capture some key factors affecting the usability of 2F technologies, and we encourage further, narrower user studies based on such metrics. Somewhat contrary to the common belief, we find that, overall, 2F technologies are perceived as highly usable, with no significant difference among them, not even when they are used for different motivations. Also, their perceived trustworthiness is not (negatively) correlated with their usability. Therefore, we conclude that user-centered design of 2F technologies, as well as R&D efforts in the field, should focus on the target population as well as the context by which these technologies will be used.

2 Related Work

This section reviews related work on the usability of multi- and single-factor authentication.

2.1 Usability of Multi-Factor Authentication Technologies

Previous work studied the usability-security trade-off of 2F, and concluded that security usually reduces usability. Braz et al. [6] are among the first to do so: they propose two rating scales (security and usability, respectively) and use them to compare user authentication methods, including 2F. They note that 2F increases “redundancy,” thus augmenting security but decreasing usability.

Strouble et al. [27] analyze the effects of implementing 2F on productivity. They focus on the “Common Access Card” (CaC), a combined smart card/photo ID card used (at that time) by US Department of Defense (DoD) employees. They present the result of a 40-item survey administered to 313 US Air Force employees, and report that users stopped checking emails at home (due to the unavailability of card readers) and that many employees accidentally left their card in the reader. Authors estimate that the DoD spent about \$10.4M on time lost (e.g., when employees left the base without their card and were unable to re-enter) and conclude that the CaC increased security at the expense of productivity.

Gunson et al. [14] focus on the usability of single and two-factor authentication in automated telephone banking. They present a survey involving 62 users of telephone banking, where participants were asked to rate their experience using a proposed set of 22 usability-related questions. According to their analysis, 2F was perceived to be more secure, but again less usable, than simple passwords and PINs.

Weir et al. [30] compare usability of push-button tokens, card activated tokens, and card & pin activated tokens. They consider usability in terms of efficiency, and measure the authentication time of users, as well as in terms of satisfaction, by asking users to rate their experience using a proposed set of 30 questions. In addition to usability, they measure quality, convenience and security. They show that users value convenience and usability over security, and thus quality and usability are sacrificed when increasing layers of security are required.

Somewhat closer to our work is another study by Weir et al. [31], that analyzes the usability of passwords and two methods of 2F: codes generated by token and PINs received via SMS. They perform a lab study where 141 participants were asked to report on the usability of the three technologies using 30 proposed questions. Authors conclude that familiarity with a technology (rather than perceived usability) impacted user willingness to use a given authentication technology. Their results show that users perceive the 1-factor method (with which the average user had most experience) as being the most secure and most convenient option.

Our work differs from [30, 31] in several aspects. We compare a larger diversity of 2F technologies (security tokens, codes received via SMS/email, and dedicated apps). Also, we do not study the trade-off between security and usability, rather, we provide a comparative study among different technologies, aiming to understand how each 2F technology

performs compared to others. Specifically, we study the relation between 2F technologies and the contexts in which they are used, as well as the motivation driving the users to adopt them. This was motivated by previous work, e.g., the studies by Goffman [13] and Nissenbaum [21] who showed that human behavior often significantly differs based on context. Finally, we consider a larger pool of participants, measure an extensive list of factors (inspired from questions proposed by [6, 14, 30, 31]), and conduct an exploratory factor analysis to determine key factors that affect usability of 2F.

2.2 Usability of Single Factor Technologies

The analysis of the usability of 2F technologies is naturally linked to that of password-based authentication and, in general, of security software.

One of the first studies in *usable security* [10] is the 1999 article by Whitten et al. [33], which presents a cognitive walkthrough of PGP.² Whitten et al. highlight properties unique to the design of usable security systems: (i) for most users, security is not a primary task, and (ii) security software is often only as strong as the weakest link, i.e., the user. Also in 1999, Adams et al. [1] show that users feel under attack by “capricious” password policies. Good security usually requires long and hard-to-remember passwords, frequent password changes, and different passwords across different services. Unfortunately, this ultimately pushes the user to find the simplest password that barely complies with requirements. Similar to [33], Adams et al. conclude that security is not a primary task: users do not sit at a computer to “be secure”, rather, they use it to complete tasks such as banking, shopping, and socializing. Thus, users tend to minimize efforts necessary to ensure security, and avoid distraction from their primary tasks.

Florencio et al. [12] show that users memorize an average of 6.5 passwords, and each password is shared across an average of 3.9 sites. Bardram et al. [3] discuss burdens on nursing staff created by hard-to-remember passwords in conjunction with frequent logouts required by healthcare security standards, such as the Health Insurance Portability and Accountability Act (HIPAA). Inglesant et al. [16] analyze “password diaries” maintained by study participants, recording the times they authenticated via passwords. Authors find that frequent password changes are a burden, that users do not change passwords unless forced to, and that it is difficult for them to create memorable, secure passwords adhering to the policy.

A possible approach for improving password security involves “password meters,” which provide visual clues and help users create memorable yet complex passwords complying with policies. Komanduri et al. [20] show that complex password policies can sometimes *decrease* average password entropy, and that a 16-character with no additional requirements provided the highest average entropy per password.

²Pretty Good Privacy (PGP) [35] is a security software for encrypting and signing emails, text, and files.

Along similar lines, Egelman et al. [11] find that for “important” accounts, a password meter successfully helps increase entropy.

Rather than increasing the entropy of user-generated passwords, another strategy is to automatically generate and store high-entropy passwords in a “password manager.” Chiasson et al. [8] compare the usability of two password managers (PwdHash and Password Multiplier), pointing to a few usability issues in both implementations. They also find that users were often uncomfortable “relinquishing control” to password managers. Karole et al. [18] study the usability of three password managers (LastPass, KeePassMobile, and Roboform2Go), with a focus on mobile phone users. They conclude that users preferred portable, stand-alone managers over cloud-based ones, despite the better usability of the latter, as they were not comfortable giving control of their passwords to an online entity.

Finally, Bonneau et al. [5] independently evaluate – i.e., without conducting any user study – different authentication schemes including: plain passwords, OpenID [22], security tokens, phone-based tokens, etc. They use a set of 25 subjective factors: 8 measuring usability, 6 measuring deployability, and 11 measuring security. Although they do not conduct any user study, authors conclude that: (i) no existing authentication scheme does best in all metrics, and (ii) technologies that one could classify as 2F do better than passwords in security but worse in usability.

Although not directly related to our 2F study, we anticipate that, in our exploratory factor analysis, we will use some of the metrics introduced in [5] and [18] (in the context of password replacements and password managers, respectively).

3 Study 1: Preliminary Interviews with 2F Users

Our first step is to determine broad trends and attitudes of 2F users: we aim to obtain a general understanding of popular 2F technologies, the context in which these technologies are used, and why they are adopted. To this end, we conducted a qualitative study involving 9 participants, which helped us drive the design of a larger quantitative study (detailed in next section). Both our initial study as well as its follow up were approved by an Institutional Review Board (details are omitted to preserve submission’s anonymity).

3.1 Methodology

We recruited participants by posting to local mailing lists and social media (Google+ and Facebook), announcing paid interviews for a user study on security and authentication technologies. Interested users were invited to complete an online pre-screening survey to assess eligibility to participate. We collected basic demographic information such as age, gender, education level, familiarity with Computer Se-

curity, and asked potential participants whether or not they had previously used 2F. A total of 29 people completed the pre-screening survey and we selected 9 participants with a wide range of ages (21 to 49), genders (5 men, 4 women), and educational background (ranging from high school to Ph.D. degrees). 5/9 users reported to have a background in Computer Security.

Interview Protocol. We interviewed users in one-on-one meetings, either face to face, or via Skype. Before each interview, users were given a consent form, indicating the interview procedure and data confidentiality. Each participant was compensated with a \$10 Amazon Gift Card.

We started the interviews by reading from a list of 2F technologies, asking participants if they had used them:

- PIN from a paper/card (one-time PIN)
- A digital certificate
- An RSA token code
- A Verisign token code
- A Paypal token code
- Google Authenticator
- A PIN received by SMS/email
- A USB token
- A smartcard
- A previously selected picture

To assess users' understanding and familiarity with 2F, we let them provide a brief description of two-factor authentication, and explain the difference with password-based authentication. (Obviously, we did not provide users with a 2F definition prior to this question, nor mention that the study was about 2F).

Then, we asked participants *why* they used 2F and why they thought other people would; this helped us understand the motivation and the context in which they used 2F. Users were also asked to recall the last time they had used any 2F technology and report any encountered issues and whether or not they wanted to change the technology (and, if so, how). If users had used multiple technologies, we also asked to compare them, and this helped us understand how participants use and perceive 2F technologies.

3.2 Findings

We found that the 2F technologies that study participants used the most included: codes generated by a *security token*, received via *SMS or email*, and codes generated by a dedicated *smartphone app*, entered along with username and password.

Participants used 2F technology in three contexts: *work* (e.g., to log into their company's VPN), *personal* (e.g., to protect a social networking account), or *financial* (e.g., to gain access to online banking).

Study participants used 2F because they either were *forced to*, *wanted to*, or *had an incentive*. Most users who adopted security tokens did so because an employer or bank had forced them. Some were unhappy about this: A participant

mentioned 2F was not "worth spending 5 minutes for simple \$1.99 purchases". Two participants (customers of different banks) reported adopting 2F in order to "obtain higher limits on online banking transactions." Other users used 2F to "avoid getting hacked."

Some users of tokens complained that it was annoying to have to remember to carry security tokens. One user recommended to "store the token in the laptop bag" to avoid this issue. Some users experienced delays from SMS-based codes, and were "annoyed, especially when paying for incoming texts." One user pointed out that (s)he "preferred text messages", since (s)he "did not have a smartphone." Others preferred not to use security tokens as they "can be lost." Some participants preferred tokens as they are easier to use compared to mobile applications, where one has to "look down to unlock screen, find app, open app, and read the code."

As mentioned above, results from Study 1 drove the design of a larger (survey-based) user study—presented in next section—providing a quantitative analysis of 2F technologies' usability.

4 Study 2: Quantitative Analysis of 2F Users' Preferences

Our second (and main) study consists of a quantitative analysis of 2F users' preferences. Inspired by the results of Study 1, we designed and conducted a survey involving 219 2F users, recruited on Mechanical Turk.

4.1 Methodology and Study Design

4.1.1 Participants' Recruitment and Demographics

We initially recruited 268 Mechanical Turk users. 13 of them were not considered eligible to participate as they had not used any 2F technology. Additionally, 36 users abandoned the survey prior to completion. The remaining 219 Mechanical Turk users were asked to complete an online survey about 2F technologies and received \$2.00 for approximately 30 minutes of survey taking.

Previous research showed that Mechanical Turk users are a valid alternative to traditional human subject pools in social sciences. For example, Jacobson [17] compared results of a study conducted on Mechanical Turk with results conducted by an independent survey company, and found that both results were statistically indistinguishable. Furthermore, Mechanical Turkers are often more diverse in terms of age, income, education level, and geographic location than the traditional pool for social science experiments [15].

Nonetheless, as pointed out by Kittur et al. [19], Mechanical Turk users often try to cheat at tasks. Therefore, we put in place several anti-fraud and sanity-check mechanisms. We screened potential participants by asking whether they had used 2F, and presented a list of examples: security tokens,

Gender	
Male	61.6%
Female	38.4%
Age	
18–24	22.4 %
25–34	48.4 %
35–44	17.8 %
45–54	5.4 %
55–65	5.4 %
65+	0.5 %
Income	
Less than \$10,000	15.5 %
\$10,000 – \$20,00	14.6 %
\$20,001 – \$35,000	25.5 %
\$35,001 – \$50,000	18.3 %
\$50,000 – \$75,000	18.7 %
\$75,001 – \$90,000	3.6 %
\$90,000 – \$120,000	2.7 %
\$120,001 – \$200,000	0.9 %
Education	
Less than high school	0.46 %
Some college	32 %
Undergrad	37.4 %
Some grad school	3.1 %
Master’s degree	5.9 %
PhD	0.9 %
Familiar with Computer Science?	
Yes	22.8%
No	77.2%
Familiar with Computer Security?	
Yes	5.4%
No	94.6%

Table 1: Participant demographics for Study 2 (Total n = 219).

codes received via SMS/email, and dedicated smartphone apps. Users who reported to have never used any of these technologies were told that they were not eligible to participate in our survey, and blocked from proceeding further or going back to change their answer via a cookie-based mechanism. Also note the Mechanical Turk task announcement did not state that users were required to have used 2F and merely presented it alongside other basic demographics such as age and gender. By hiding the purpose of our task, we were able to avoid incentivizing Turkers to lie.

Furthermore, the survey included several sanity-check questions, such as simple math questions (in the form of a Likert question) to verify that participants were paying attention, and not answering randomly. Similarly, we introduced some contrasting Likert questions (e.g., “I enjoyed using the technology” and “I did not enjoy using the technology”) and verified that answers were consistent. Users who did not answer correctly all sanity checks were to be dis-

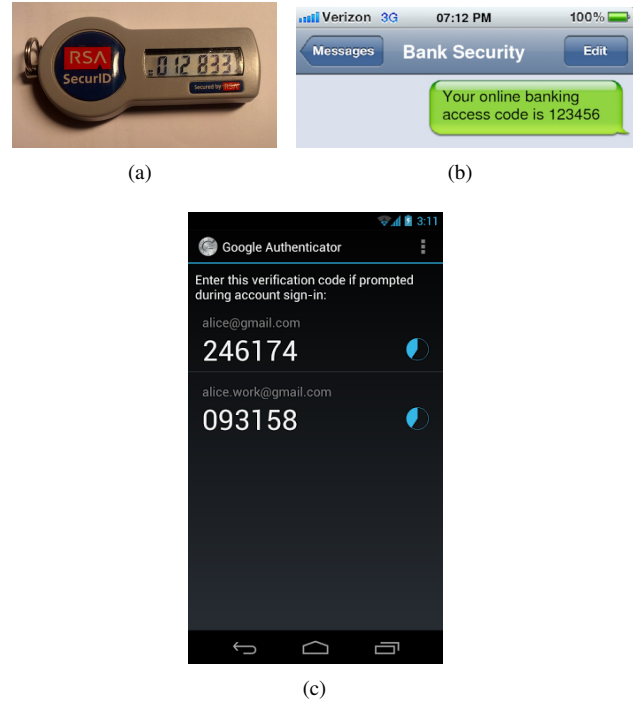


Figure 1: Examples of 2F technologies: (a) codes generated by a security token, (b) codes received via SMS, (c) codes generated by a dedicated smartphone app.

carded from the analysis (but still compensated), however, all users who made it past our initial screening actually answered the sanity-checks correctly. Finally, analysis of the time spent by each survey participant showed completion times in line with those of test runs done by experimenters.

The demographics of the 219 study participants are reported in Table 1. Our population included 135 (61.6%) males and 84 females (38.4%). 50/219 (22.8%) users reported a background in computer science, and 12/219 (5.4%) users reported a background in computer security. Education levels ranged from high school diploma to PhD degrees. Ages ranged from 18 to 66, with an average age of 32 and a standard deviation of 10.2.

4.1.2 Technologies, Context, and Motivation

The first question in the survey asked users if they had used any of the following 2F technologies (for each of them, we displayed an example picture):

Token: Standalone pieces of hardware which display a code, Figure 1(a).

Email/SMS: A code received via email or SMS (aka text message), Figure 1(b).

App: Codes delivered via an app running on a smartphone or other portable electronic device, such as an iPad or Android tablet, Figure 1(c).

Next, the survey branched depending on how many and what technology/technologies had been selected. Specifically, *users were asked to answer the same set of questions for each technology they had used.*

One of our main objectives was to measure and compare in which context and with what motivation users were exposed to 2F technologies. Specifically, for each technology we asked users in which of the following *context(s)* they used the technology:

Financial: While doing online banking or other financial transactions (e.g., bill payment, checking credit card balance, doing taxes).

Work: While performing work duties (e.g. logging in company VPN).

Personal: While accessing a personal account not used for work or finance (e.g., Facebook, Twitter, Google, etc.).

Also, we asked users *why* they had been using 2F. Possible motivations included:

Voluntary: The participant voluntarily adopted 2F.

Incentive: The participant got an incentive to adopt 2F (e.g., extra privileges/functionality, such as increased bank transfer limits).

Forced: The participant had no choice (e.g., employer policy forcing adoption).

4.1.3 System Usability Score and Other Likert Questions

For each employed 2F technology, participants were asked to rank the usability of the technology using 10 Likert questions from System Usability Scale (SUS) [7]. Previous research has shown SUS is a fairly accurate measure of usability [2].

Note that, in order to be consistent with other Likert questions in our survey, we modified the SUS questionnaire to include a 7-point range, rather than the more common 5-point range, with 1 being “Strongly Disagree” and 7 being “Strongly Agree”. Conversion from a 7-point Likert range to a 5-point Likert range is simply done via normalization.

Next, *for each employed 2F technology*, participants, where asked a series of 7-point Likert questions (with 1 being “Strongly Disagree” and 7 being “Strongly Agree”) about the following statements:

- Convenient: I thought (technology) was convenient.
- Quick: Using (technology) was quick.
- Enjoy: I enjoyed using (technology).
- Reuse: I would be happy to use (technology) again.
- Helpful: I found using (technology) helpful.
- No Enjoy: I did not enjoy using (technology).
- User Friendly: I found (technology) technology user friendly.
- Need Instructions: I needed instructions to use (technology).
- Concentrate: I had to concentrate when using (technology).
- Stressful: Using (technology) was stressful.

Group	2F Technologies	# of Participants
1	Token	11
2	Email/SMS	77
3	App	7
4	Token & Email/SMS	29
5	Token & App	3
6	Email/SMS & App	50
7	Token, Email/SMS & App	41
Total		219

Table 2: Usage of 2F technologies among survey participants.

- Match: (technology) did not match my expectations regarding the steps I had to follow to use it.
- Frustrating: Using (technology) was frustrating.
- Trust: I found using (technology) trustworthy.
- Secure: How secure did you feel to authenticate using (technology) instead of just username & password? (1: “Not at All Secure” 7: “Very secure”)
- Easy: Knowing how to get the code from (technology) was easy.

The above questions are inspired from metrics used in previous work (e.g., [5, 18]) and findings from our Study 1. They are meant to be extensive and measure factors beyond the System Usability Score, such as trustworthiness, convenience, ease of use, reuse, enjoyment, concentration, portability, etc.

4.2 Study 2 Results

We first analyze how 2F technologies are used by investigating the relation between independent factors such as context, motivation, technologies, and gender. We then provide an exploratory factor analysis about users’ perception of 2F technologies (Likert questions), aiming to understand which factors are best to capture the usability of 2F. We then provide a comparative analysis of the usability of 2F technologies using those factors, and conclude with a discussion of our findings and highlighting some issues with 2F.

4.2.1 Use of 2-Factor

Recall, from our study design, that participants were asked to identify the different 2F technologies they use, in which context, and why. Almost half of the participants (43%) used only one technology, while 37% used two, and 20% three technologies. Table 2 summarizes the use of the three 2F technologies among the 219 participants. We observe that “Email/SMS” (i.e., one-time codes received via SMS or email) is the most used technology as 89.95% (197/219) used it as a second factor. Also, 45.20% (99/219) of participants used “App” (i.e., codes generated by a dedicated smartphone app, such as Google Authenticator). “Token” (i.e., codes generated by a hardware/security token) is the least common technology, only used by 24.20% (53/219).

It is interesting to notice that App, despite being the most

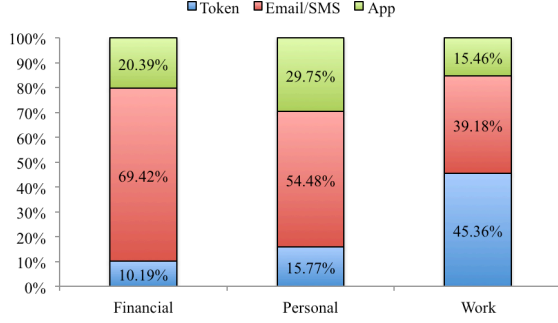


Figure 2: Distribution of the use of 2F technologies across contexts.

recent technology, has a higher adoption rate than Token, one of the oldest technology. This evolution might be related with the fast-increasing number of users owning smartphones, which can serve as a second-factor device that is always with the user.

Different Technologies in Different Contexts. The three 2F technologies are used differently depending on context (see Figure 2). In the financial context, Email/SMS is the most popular 2F (69.42%), followed by App (20.39%) and Token (10.19%). In the personal context, Email/SMS is also the most popular (54.48%), followed by App (29.75%) and Token (15.77%). In the work context, Token is the most popular (45.36%), followed by Email/SMS (39.18%) and App (15.46%). A χ^2 -test shows that differences are significant ($\chi^2(4, N = 582) = 65.18, p < 0.0001$).

It is relatively unsurprising that Token is most popular in the work context—an environment with high inertia—while it is noticeable that many users adopt tokens in the personal context. The analysis of open-ended questions seem to show that online gaming is the main field of adoption for Token in the personal context.

Different Motivations for Different Technologies. We find that few participants are incentivized to use 2F – see Figure 3. Only 19.73% of Token users, 11.65% of Email/SMS users and 9.25% of App users are incentivized. Actually, 44.90% of Token users were forced, while 53.18% of App were voluntary. A χ^2 -test shows that differences are significant ($\chi^2(4, N = 775) = 14.68, p < 0.001$).

Different Motivations in Different Contexts. We find that in the work context, 60.84% of participants were forced to use 2F, versus 27.97% of participants using 2F voluntarily. In the personal context, more than half participants (51.26%) use 2F voluntarily and 34.73% are forced to. In the financial context, about 45.45% of participants use 2F voluntarily and 42.91% are forced to. Distributions are plotted in Figure 4. A χ^2 -test shows that differences are significant ($\chi^2(4, N = 775) = 29.76, p < 0.0001$).

This result is expected, as users tend to be forced to use 2F at work, and tend to use it voluntarily (opt-in) for personal

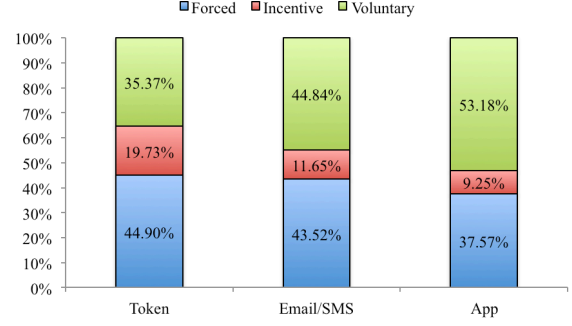


Figure 3: Distribution of the motivation across 2F technologies.

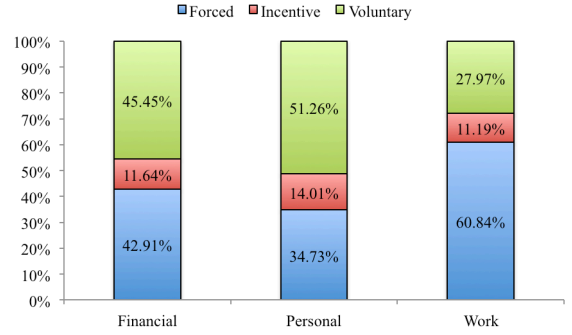


Figure 4: Distribution of the motivation across contexts.

use. In the financial context, the distribution is even.

Gender Differences. While there is no gender difference in terms of adoption rate for Token and Email/SMS, male users adopt App-based 2F more than female users – see Table 3. The χ^2 -test shows that the difference is significant ($\chi^2(1, N = 219) = 29.76, p < 0.05$).

2F for Online Gaming. As mentioned earlier, we also asked participants to list services and websites for which they used 2F. Surprisingly, we find that in the personal context, in addition to personal email, document sharing, and social networking sites, participants also used 2F for online gaming, e.g., on Battle.net, Diablo 3, World of Warcraft, Blizzard Entertainment, and swtor.com.

4.2.2 An Exploratory Factor Analysis

While SUS is a generic usability measure, we argue that 2F technologies rely on a unique combination of hardware and software that SUS may fail to capture. Following this shortcoming, previous work [5, 6, 14, 18, 30, 31] considered a series of questions and parameters to evaluate the usability of 2F schemes. In order to obtain key elements central to the understanding of the usability of 2F, we perform an exploratory factor analysis (see aforementioned 15 Likert questions).

We factor-analyze our questions using Principal Component Analysis (PCA) with Varimax (orthogonal) rotation. Items with loadings < 0.4 are excluded. The analysis yields

	Loadings			
	Factor 1: Ease of Use	Factor 2: Cognitive Efforts	Factor 3: Trust	Communality
Convenient	0.91	0.05	-0.02	0.77
Quick	0.84	-0.12	-0.15	0.67
Enjoy	0.77	0.15	0.12	0.63
Reuse	0.75	0.04	0.19	0.75
Helpful	0.72	0.02	0.17	0.69
No Enjoy	-0.52	0.22	-0.16	0.55
User Friendly	0.42	-0.19	0.37	0.74
Need Instructions	0.15	0.80	-0.12	0.60
Concentrate	0.03	0.64	0.14	0.38
Stressful	-0.41	0.51	0.01	0.59
Match	-0.30	0.42	-0.15	0.47
Frustrating	-0.47	0.47	0.00	0.63
Trust	0.08	-0.04	0.80	0.74
Secure	-0.02	0.03	0.82	0.82
Easy	0.27	-0.28	0.31	0.44
Eigenvalues	7.52	1.78	1.03	
% of Variance	32	15	14	
Total Variance		61%		

Table 4: Factor Analysis Table.

	Female	Male
App Users	31	71
Non-App	53	64

Table 3: Distribution of gender across 2F App technology.

three factors explaining a total of 61% of the variance for the entire set of variables. These factors are independent of each other (i.e., they are not correlated).

Factor 1 is labeled ***Ease of Use*** (EaseUse for short) due to the high loadings by following items: Quick, enjoy, user friendly, convenient, easy reuse, helpful, and convenient. This first factor explains 32% of the variance.

The second derived factor is labeled ***Cognitive Efforts*** (CogEfforts for short). This factor is labeled as such due to the high loadings by following factors: Frustrating, stressful, match, need instructions, and concentration. The variance explained by this factor is 15%.

The third derived factor is labeled ***Trustworthiness***. This factor is labeled as such due to the high loadings by following factors: Secure and trust. The variance explained by this factor is 14% (Table 4).

The communalities of the variables included are rather low overall, with one variable having a small amount of variance (Concentrate, 38%) in common with the other variables in the analysis. This may indicate that the variables chosen for this analysis are only weakly related with each other. However, the KMO test indicates that the set of variables are at least adequately related for factor analysis.

In conclusion, we have identified three clear factors among participants: ease of use, required cognitive efforts, and trustworthiness.

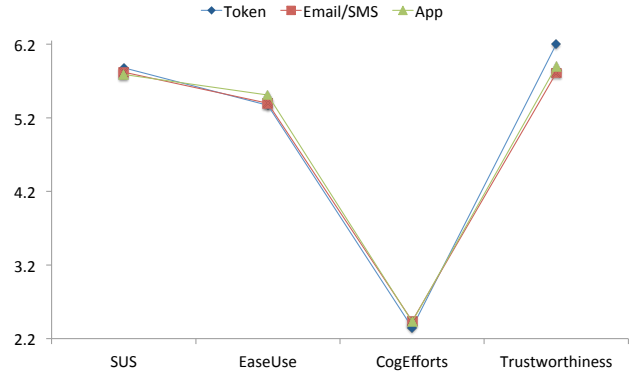


Figure 5: Overview of usability measures of different 2F technologies.

4.2.3 Overview of Usability Measures

In Figure 5, we show the average usability measures for different 2F technologies. We obtain similar ratings for different technologies. The average SUS score is around 5.8, EaseUse is about 5.4, CogEfforts is 2.4, and Trustworthiness is around 6. We observe that SUS is high for all 2F technologies: converting the SUS score to a percentage scale, overall SUS is more than 80%, which is considered “Grade A” usability [24]. In addition, SUS is correlated with EaseUse ($r = 0.8$). Next, we look into factors that influence 2F usability measures.

4.2.4 Comparison among Different 2F Technologies

We now compare the usability of different 2F technologies, taking into consideration the context in which they are

used and the characteristics of the individuals who used them, including age, gender and whether they have a computer science background (“CS_Back”). Some participants only use one of the 2F technologies and others use more than one. To compare the usability of different technologies, we split participants into 7 subgroups (Table 2) and performed analysis on each of the subgroups. Since there are not enough participants in Groups 1, 3, and 5 (Table 2), these groups were not analyzed. For usability measures, we use the three newly discovered factors (introduced above): “EaseUse” ($\alpha = 0.92$), “CogEffort” ($\alpha = 0.74$), and “Trustworthiness” ($\alpha = 0.81$).

Email/SMS Users. 77 participants only use Email/SMS as 2F technology, with 13 of them having a CS background. To compare the usability measures for participants who only use Emails/SMS (Group 2), we ran a MANOVA with one between factor computer science background (CS_Back vs non CS_Back), and age, gender and context as covariates. The dependent variables were the three usability measures. Using Pillai’s trace, CS_Back ($V = 0.07$, $F(3, 124) = 3.04$, $p = 0.02$) was a significant factor. Age, gender and context were not significant.

Further analysis shows that participants without a computer science background (EaseUse $\mu = 5.67$, $\sigma = 1.03$) find Email/SMS to be more usable than participants with a computer science background (EaseUse $\mu = 5.07$, $\sigma = 1.88$, $t(35.94) = 2.94$, $p = 0.006$).

Token & Email/SMS Users. 29 participants used both token and email/SMS 2F technologies, with 7 having a computer science background. To compare the usability measures for participants who use both Emails/SMS and app (Group 4), we ran a one way (Technology: Token vs. Email/SMS) within subject MANOVA, with age, gender and context as covariates. The CS_Back was not included in the analysis because there were not enough participants with CS_Back. No main effect of technology was found. Age was a significant covariate ($V = 0.13$, $F(3, 63) = 3.12$, $p = 0.03$).

Further analysis shows that elder people (above average age ($\mu = 33$, $N = 12$) in Group 4) (CogEfforts $\mu = 3.04$, $\sigma = 1.29$) find more cognitive efforts were needed to use 2F technology than younger people ($N=17$) (CogEfforts $\mu = 2.19$, $\sigma = 0.98$, $t(54.24) = 3.03$, $p = 0.004$).

Email/SMS & App Users. 50 participants used both Email/SMS and App (Group 6), with only 5 having a computer science background. To compare the usability measures for participants who use both Emails/SMS and app (Group 6), we ran a one way (Technology: Email/SMS vs. App) within subject MANOVA, with age, gender and context as covariates. The CS_Back was not included in the MANOVA because there were not enough participants. Similar to results in Group 4, no main effect of technology was found. Age was a significant covariate ($V = 0.13$, $F(3, 63) = 3.12$, $p = 0.03$).

Further analysis shows that elder people (above average age ($\mu = 32$, $N = 25$) in Group 6 (CogEfforts

$\mu = 2.81$, $\sigma = 1.25$ marginally significant) find more cognitive efforts were needed to use 2F technology than younger people ($N = 25$) (CogEfforts $\mu = 2.51$, $\sigma = 1.05$, $t(153.50) = 1.68$, $p = 0.09$, marginally significant). Elder people find 2F technology less usable (EaseUse $\mu = 5.16$, $\sigma = 1.03$) and less trustworthy (Trustworthiness $\mu = 5.58$, $\sigma = 0.96$) than younger people (EaseUse $\mu = 5.45$, $\sigma = 1.03$, $t(164.61) = -1.78$, $p = 0.07$, Trustworthiness $\mu = 5.95$, $\sigma = 0.87$, $t(158.41) = -2.61$, $p = 0.01$).

Token, Email/SMS & App Users. 41 participants used all token, email/SMS and App, with 17 having a CS background. To compare the usability measures for participants in this group, we ran a 3(Technology: Token, Email/SMS vs. App) x 2(CS_Back vs non CS_Back) MANOVA, with Technology as a within subject variable and CS_Back as a between subject variable, and age, gender and context as covariates. The main effects of technology and CS.back were not significant. Gender is a significant factor ($V = 0.12$, $F(3, 168) = 7.44$, $p = 0.0001$).

Female users (CogEfforts $N = 12$, $\mu = 2.58$, $\sigma = 1.11$) found more cognitive efforts were required than male users (CogEfforts $N = 29$, $\mu = 2.01$, $\sigma = 0.89$, $t(56.23) = 2.52$, $p = 0.01$).

4.2.5 Analysis of Open-Ended Questions

For each 2F technology, we asked users to answer a few open-ended questions about the services/websites where they used 2F and they issues they encountered. As mentioned earlier, security tokens tend to be used for work, finance, and personal websites. Interestingly, users often rely on tokens to protect their online gaming accounts: the fear of losing their gaming profile is high enough for users to adopt 2F. Users complain that the authentication process is often prone to failure (“The authentication to the server was down.”), is time sensitive (“Sometimes, during the code rollover, you’d end up with a mismatch and have to start the whole process over”), and that problem resolution is complicated (“If I made three mistakes entering my code, I had to call the state help desk to have my PIN reset”).

Email/SMS have, overall, a high variety of use cases, but were frequently used with banks as well as with Facebook, Google, and Paypal. People complained about specific issues with codes expiring or failing to be received, especially while traveling abroad. For instance, a number of users complain about SMS not working abroad (“Sometimes it wouldn’t send”, “My husband changed his phone number when moving to the US and had a lot of problems getting things.”, “Sometimes I am unable to receive a code if I am overseas. In that case, I have to call a toll free number or e-mail customer support to receive the code via e-mail instead of text.”), and again regarding the difficult problem resolution (“The passcode they sent me didn’t work and I had to call them to get a new one. It was very frustrating.”).

Finally, app-based 2F was often used with popular service

providers (such as, Google and Facebook), banks, gaming platforms, and emails. No specific issues were reported.

5 Discussion

Based on our quantitative study of 2F technologies, we can draw some conclusions and highlight the need for further research on 2F.

We noticed that technologies are adopted at different rates depending on *contexts* and *motivations*. Specifically, in the work environment, codes generated by security tokens constitute the most used second authentication factor. Whereas, codes received via email or SMS are most popular in the financial and personal contexts. Also, few users receive incentives to adopt 2F, while many utilize security tokens because they are forced to, or decide to opt-in to use dedicated smartphone apps.

These observations, along with answers to open-ended questions, indicate that: (1) financial institutions introduced 2F as mandated by policy [9] but did so while minimizing the “effort” imposed on customers,³ (2) enterprises rely on (mostly proprietary) security tokens (e.g., RSA/Verisign tokens) for authentication to corporate networks, and (3) smartphone apps (e.g., Google Authenticator) are used by customers who opt-in to 2F with online services providers, such as, Google, Dropbox, or Facebook.

Another relevant finding is that users’ perception of trustworthiness is not negatively correlated with ease of use and required cognitive efforts. This seems to contrast with results of prior studies [6, 14]. We find that 2F technologies perceived as more trustworthy are not necessarily less usable. One possible explanation is that prior work mostly compared 2F with passwords, whereas, we compare 2F technologies to each other.

Our comparative analysis is essential to develop an understanding of which 2F technologies users prefer. Indeed, in many cases, password-based authentication is not an option (e.g., for corporate VPN access, or for some financial services), and thus more usable 2F technologies in that context should be favored. Similarly, when users have the choice to opt-in, adoption rates will likely depend on 2F usability. A few companies (e.g., Google, Duo Security, PassBan, Authy) are researching ways to improve 2F usability and adoption rates. In our study, we find that, contrary to common assumptions, neither motivation nor the type of 2F technology significantly affect our usability measures. Instead, age, gender, and Computer Science background do. This surprising result suggests that innovation in the scope of 2F and user-centered design of 2F technologies should really focus on specific target populations, as well as the context in which these technologies will be used.

³For instance, some financial institutions (e.g., Chase and Bank of America) only require the second factor to be entered if a user is authenticating from an “unrecognized” device.

6 Conclusion

This paper presented the first large-scale comparative study of two-factor authentication (2F) technologies. First, we reported on a preliminary, interview-based study (involving 9 participants), intended to identify popular 2F technologies as well as how they are used, when, where, and why. Next, we designed and administered an online survey to 219 Mechanical Turk users, aiming to measure the usability of a few popular 2F technologies: one-time codes generated by security tokens, one-time PINs received via SMS or email, and dedicated smartphone apps. We also recorded contexts and motivations, and study their impact on perceived usability of different 2F technologies.

We presented an exploratory factor analysis to evaluate a series of parameters, including some suggested by previous work, to evaluate the usability of 2F, and show that ease of use, trustworthiness, and required cognitive effort are three key aspects defining 2F usability. Finally, we showed that differences among the usage of 2F depend on individual characteristics of people, rather than the actual technologies or contexts. We considered a few characteristics, such as age, gender and computer science background, and obtained a few insights into user preferences. Our analysis suggests that user-centered design of 2F technologies, as well as R&D efforts in the field, should focus on the target population as well as the context by which these technologies will be used.

Our results provide a few insights for future work. Besides encouraging follow-up research building on our factor analysis, we also plan to conduct narrower studies for some specific technologies and target audiences.

References

- [1] A. Adams and M. A. Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.
- [2] A. Bangor, P. T. Kortum, and J. T. Miller. An empirical evaluation of the system usability scale. *Intl. Journal of Human-Computer Interaction*, 24(6):574–594, 2008.
- [3] J. E. Bardram. The trouble with login: on usability and computer security in ubiquitous computing. *Personal and Ubiquitous Computing*, 9(6):357–367, 2005.
- [4] L. Bauer, L. F. Cranor, M. K. Reiter, and K. Vaniea. Lessons learned from the deployment of a smartphone-based access-control system. In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, pages 64–75, 2007.
- [5] J. Bonneau, C. Herley, P. C. van Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 553–567, 2012.
- [6] C. Braz and J.-M. Robert. Security and usability: the case of the user authentication methods. In *Proceedings of the 18th International Conference of the Association Francophone d’Interaction Homme-Machine*, pages 199–203, 2006.

- [7] J. Brooke. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189:194, 1996.
- [8] S. Chiasson, P. C. van Oorschot, and R. Biddle. A usability study and critique of two password managers. In *USENIX Security Symposium*, pages 1–16, 2006.
- [9] Council, Federal Financial Institutions Examination. Authentication in an internet banking environment. http://www.ffiec.gov/pdf/authentication_guidance.pdf, 2005.
- [10] L. F. Cranor and S. Garfinkel. *Security and usability: Designing secure systems that people can use*. 2007.
- [11] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley. Does my password go up to eleven?: the impact of password meters on password selection. In *CHI*, 2013.
- [12] D. Florencio and C. Herley. A large-scale study of web password habits. In *WWW*, pages 657–666, 2007.
- [13] E. Goffman. *The Presentation of Self in Everyday Life*. 1959.
- [14] N. Gunson, D. Marshall, H. Morton, and M. Jack. User perceptions of security and usability of single-factor and two-factor authentication in automated telephone banking. *Computers & Security*, 30(4):208–220, 2011.
- [15] J. Henrich, S. J. Heine, A. Norenzayan, et al. The weirdest people in the world. *Behavioral and Brain Sciences*, 33(2-3):61–83, 2010.
- [16] P. G. Inglesant and M. A. Sasse. The true cost of unusable password policies: password use in the wild. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 383–392, 2010.
- [17] M. Jakobsson. Experimenting on mechanical turk: 5 how tos. *ITWorld*, September, 3:2009–2009, 2009.
- [18] A. Karole, N. Saxena, and N. Christin. A comparative usability evaluation of traditional password managers. In *ICISC*, 2011.
- [19] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456, 2008.
- [20] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman. Of passwords and people: measuring the effect of password-composition policies. In *CHI*, 2011.
- [21] H. Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79:119, 2004.
- [22] D. Recordon and D. Reed. OpenID 2.0: a platform for user-centric identity management. In *ACM Workshop on Digital Identity Management*, 2006. <http://openid.net>.
- [23] A. P. Sabzevar and A. Stavrou. Universal multi-factor authentication using graphical passwords. In *SITIS*, pages 625–632, 2008.
- [24] J. Sauro and J. R. Lewis. *Quantifying the user experience: Practical statistics for user research*. Kaufman, 2012.
- [25] B. Schneier. Two-factor authentication: too little, too late. *Commun. ACM*, 48(4):136, 2005.
- [26] B. Schneier. iPhone Fingerprint Authentication. https://www.schneier.com/blog/archives/2013/09/iphone_fingerpr.html, 2013.
- [27] D. D. Strouble, G. Schechtman, and A. S. Alsop. Productivity and usability effects of using a two-factor security system. *Proceedings of SAIS*, pages 196–201, 2009.
- [28] A. Vance. If your password is 123456, just make it hackme. *New York Times*, <http://nyti.ms/7GmDh2>, 2010.
- [29] E. von Zeszschwitz, A. De Luca, and H. Hussmann. Survival of the Shortest: A Retrospective Analysis of Influencing Factors on Password Composition. In *INTERACT*, pages 460–467, 2013.
- [30] C. S. Weir, G. Douglas, M. Carruthers, and M. Jack. User perceptions of security, convenience and usability for ebanking authentication tokens. *Computers & Security*, 28(1):47–62, 2009.
- [31] C. S. Weir, G. Douglas, T. Richardson, and M. Jack. Usable security: User preferences for authentication methods in ebanking and the effects of experience. *Interacting with Computers*, 22(3):153–164, 2010.
- [32] R. Weiss and A. De Luca. Passshapes: utilizing stroke based authentication to increase password memorability. In *Nord-CHI*, pages 383–392, 2008.
- [33] A. Whitten and J. D. Tygar. Why Johnny can’t encrypt: A usability evaluation of PGP 5.0. In *Proceedings of the 8th USENIX Security Symposium*, volume 99, 1999.
- [34] J. Yan, A. Blackwell, R. Anderson, and A. Grant. Password memorability and security: Empirical results. *IEEE Security & Privacy*, 2(5), 2004.
- [35] P. R. Zimmermann. *The official PGP user’s guide*. MIT press, 1995.