2FA Might Be Secure, But It's Not Usable: A Summative Usability Assessment of Google's Two-factor Authentication (2FA) Methods

Claudia Ziegler Acemyan[1], Philip Kortum[1], Jeffrey Xiong[1, 2], and Dan S. Wallach[2]
Departments of Psychological Sciences[1] and Computer Science[2], Rice University
6100 Main Street, MS-25, Houston, Texas 77005, USA
{claudiaz, pkortum, dwallach@rice.edu}@rice.edu

Computer security experts recommend that people use two-factor authentication (2FA) on password protected systems to help keep hackers out. Providing two pieces of information to verify a person's identity adds extra security to an account. However, it is not clear if the added security and procedures impact system usability. This paper aims to answer this question by assessing per ISO 9241-11's suggested measurements the usability of Google's optional 2FA methods. We found few differences across four different 2FA methods when comparing efficiency, effectiveness and satisfaction measures—illustrating that one method is not necessarily more or less usable then another. Overall, the measures indicated that the systems' usability needed to be improved, especially with regard to the initial setup of 2FA. In conclusion, developers need to focus more attention on making 2FA easier and faster to use, especially since it is often optional for password users, yet makes accounts significantly more secure.

INTRODUCTION

Theoretically, system users should be using two-factor authentication (2FA), because passwords are imperfect. Malicious individuals can guess passwords using publicly available information, and cyber attackers can test large numbers of passwords in a short amount of time to try to access an account (ArunPrakash & Gokul, 2011). If successful, hackers can access money in banking accounts, purchase goods through online shops, access files stored in the cloud, take over social media feeds, and obtain every password stored in an individual's password manager. All of these events can result in tremendous financial loss and exposure of confidential, personal information. However, if two-factor authentication is used (i.e., an extra step is added to the login procedure), these types of events will be less likely (Brandom, 2017; Ong, 2018).

Two-factor authentication is not available on every password protected system, but 2FA is quickly growing in popularity as a mechanism to protect users against the many faults of passwords. Even if an attacker learns a user's password—perhaps because the user reused it across many different websites, so they would only have to memorize and recall one complex password (e.g., Morley, 2016)—the attacker must also have the "second factor" to login. The second authentication factor could be something that a user has—like a physical device (e.g., mobile phone, SIM, USB stick, key fob). Alternately, the second factor could be biological data unique to the user—such as a fingerprint or retina scan.

Of course, these different 2FA technologies have widely different security properties—making some 2FA methods easier to hack than others. For example, attackers have recently begun exploiting weaknesses in our telephony infrastructure to hijack SMS messaging (Zetter, 2016)—highlighting that even with 2FA implemented, login systems are still vulnerable to attacks. Despite the importance of making certain that the 2FA methods protect systems, this paper does not focus on how these technologies vary in their security strength; instead, it asks whether users are capable of configuring and using various 2FA methods correctly.

Understanding the usability of systems is critical. System usability is defined as how easy or hard a system is to use by actual users for a specified task in a particular environment (Bevan, 2009; Kortum, 2017). A system must be highly usable because it does not matter how secure a system is if people cannot use it (Acemyan, Kortum, Byrne, & Wallach, 2014; 2015). In the context of 2FA, if people are not able to set it up or login to a password protected system, then the added security is moot.

Unlike previous papers published by computer scientists that debate the theoretical superiority of specific 2FA techniques, this paper empirically identifies which 2FA methods are best based on objective, human-performance data and the usability measurements specified by ISO 9241-11: efficiency, effectiveness, and satisfaction (International Organization for Standardization, 1997). Each of these metrics provide us dispassionate mechanisms for understanding how usable a 2FA scheme might be.

For our study, we considered four different 2FA mechanisms used in production by Google. The participants in our studies were given Google accounts without 2FA and were asked to set up 2FA. We separately assessed the usability of this setup task, as well as the task of logging in with 2FA enabled. We selected Google's suite of 2FA techniques because they are widely used, and Google is clearly spending significant resources on the robustness and usability of their 2FA mechanisms. To our knowledge, a summative usability assessment of their techniques has not been published. An additional advantage of using Google's 2FA methods is that every one of our participants already had experience with Google's regular login process (undergraduate email at our institution is provided with a customized Gmail suite) thereby eliminating some participant variability.

This paper helps to close the gap in the limited, existing 2FA usability literature, which examined only the login

process, often did not directly observe participants or use ISO 9241-11 suggested measurements, and studied systems that were less widespread, and popular than applications such as Google and Facebook (Cristofare, Freudiger, & Norcie 2013; Weir, Douglas, Carruthers, & Jack, 2009; Gunson, Marshall, Morton, & Jack, 2011).

It is anticipated that the baseline usability data from this study will help researchers and developers determine if 2FA is generally usable and identify which 2FA methods are the easiest to use. This information can help system developers objectively choose 2FA techniques.

This research will also contribute to setting a gold standard for conducting usable security research. The protocol used here can be, and has been (e.g., Acemyan et al., 2014; 2015), used to study the usability of other secure systems. This is especially important in the computer science and security communities because their researchers are usually not human factors and usability experts familiar with the rigorous standards of system usability assessments. Due to this gap in knowledge, usable security research has been published that uses ad-hoc methods, measurements, and definitions that frequently lack scientific rigor.

METHODS

Participants

Participants included 27 Rice University undergraduates who were recruited through the school's human subjects pool. Students who volunteered to participate in the study were given partial course credit. Of these participants, data was analyzed for 20; 7 participants were excluded from data analysis due to significant equipment failures that prevented them from attempting their assigned tasks.

Eighteen of the participants were female and two were male. The mean age was 20.2 years old, with a standard deviation of 1.1 year and a range of 18 to 22 years. Most of the participants were either Asian (40%) or Caucasian (35%), followed by African American (10%), Hispanic (10%), and Multiracial (5%). Even though the participants in this study are young, from a highly selective academic institution, and mostly female, they are none the less real 2FA users. Moreover, this type of sample yields best-case-scenario usability results, as users like these are usually more computer savvy, highly motivated, seek out challenges, and have higher SAT, ACT, and GRE scores than the general US population—all of which are associated with higher cognitive abilities. Therefore, if participants from this type of sample find 2FA methods difficult or impossible to use, then certainly others will too (Byrne, 2007).

Ten (50%) subjects indicated they had used an android device before participating in this study, eight (40%) said they had not, and two (4%) did not provide this information. The mean number of hours participants use computers each week was 43.0 hours (s = 22.9), with a range of 15 to 100 hours. The mean self-reported computer expertise rating on a scale from one to ten, with one being a novice with little knowledge and ten being an expert computer user, was 7.1 (s = 2.1). When asked about their knowledge about password security, the mean rating was 5 (s = 2.4) on a scale of one to ten, with one being no knowledge and ten being advanced knowledge. The mean knowledge of computer security score, which used the same scale just described, was 3.5 (s = 2.4). Thirteen subjects (65%) indicated they had never heard of two-factor authentication before this study, in contrast to the 7 (35%) who had. Most participants (16, 80%) said they never previously used 2FA, while 4 (20%) had. System testing that includes novice users is not a limitation. Rather, systems should be designed for people who have never used them before to ensure that they can complete specified tasks on the system. Accordingly, usability assessments should include users who have no prior experience with the product.

Measures

Per ISO 9241-11, usability was assessed using efficiency, effectiveness, and satisfaction measurements. The setup and login tasks were assessed independently of one another.

Efficiency is defined as the amount of time it took participants to complete a task. The number of seconds it took participants to setup an assigned 2FA method was recorded. The amount of time it took to login to their Google/Gmail account using the 2FA technique was also recorded.

Effectiveness is defined as being able to complete a specified task using the system. It was measured by noting whether or not a user was able to setup 2FA and later login using the same method. Percentages of participants who were able to complete the task were reported. In case users did not successfully setup the 2FA method, all users were given the system with a correct configuration when asked to login so that each task could be accurately assessed.

Satisfaction is how satisfied users are with the usability of the 2FA technique. Satisfaction was measured using the System Usability Scale (SUS, Brooke, 1986). The SUS generates a score on a 100-point scale, and it can be thought about like a 100-point academic grading scale, e.g., 90-100 is an A, 80-89 is a B and so on (Bangor, Miller, & Kortum, 2009). Basic demographic information was also collected.

Materials

Testing materials included Google's two-step verification system available to the public during April 2017. Specifically, there were four 2FA techniques tested in this study; every subject used all four techniques (i.e., a within-subjects design). 1) SMS/Voice message entailed Google sending a one-time verification code to the participant's cell phone via text message or a phone call. 2) Google Authenticator App was an app that had to be downloaded on the phone, which then generated time-based verification codes. 3) USB Security Key was a security key inserted in the computer's USB port. 4) Google Prompt was a push notification sent directly to the user's registered mobile device. In all cases, this extra piece of information was given to Google after a user correctly entered their user name and password.

The following hardware was provided to participants to use during the study: an LG smartphone running Android, a Yubico security key (FIDO U2F), and a laptop computer

(Macbook Pro Late 2013) running Mac OS 10.12.4. Study instructions, a Google user name and password, phone number, and ordered list of 2FA methods was printed on paper. During the study, some participants elected to use their own Apple iPhone to complete the tasks; the participants did not ask the experimenter for permission and this is a limitation to the study. The System Usability Scale (SUS) and demographic survey were administered online through Survey Monkey.

Procedures

After participants completed IRB informed consent, they were provided with study instructions, an existing Google user name and password, a cell phone and its number, a security key, and a laptop. They were told they could use any means necessary to complete the tasks they would be asked to complete, and that any software required would not have been downloaded for them. When ready, participants signed into Google and set-up the first 2FA method assigned to them. After subjects indicated that they were done with this first task, subjects completed the System Usability Scale (SUS) to assess the usability of the setup process they just completed. Next, participants were asked to login to Google using the same 2FA method. This time, the experimenter set up the 2FA method for the account before the participant started the task, ensuring that any errors or failures in the set up did not affect the participant's ability to complete this second task. The experimenter also reset the computer and phone so that no information was stored on the devices. After participants said they completed the login task with their assigned 2FA method, they completed another SUS, keeping in mind the task they just performed on the system. This process continued until participants had used all four 2FA methods (each of the participants was randomly assigned a different ordering of the tasks to control for ordering effects; all orders were used). Next, participants completed a final study survey. Last, participants were debriefed, paid, and thanked for their time.

RESULTS

2FA Setup

Efficiency. The mean amount of time it took users to setup a 2FA system was 314 seconds, with a standard deviation of 114s and a range of 174 to 555s. As can be seen in Figure 1, there is insufficient evidence to support differences in setup times across systems, $F(1.99, 37.71)=2.42$, $MSE = 40{,}831.06$, $p=.103$, $\eta_p^2=.11$ (Greenhouse-Geisser corrected).

Effectiveness. The mean percentage of users who successfully completed the 2FA setup was 68%, with a range of 55% to 75%. As shown in Figure 2, there was not a statistically reliable effect for participants being able to setup some 2FA methods more often than others, $F(3, 57)=.839$, $MSE=.22$, $p=.478$, $\eta_p^2=.04$.

Satisfaction. For satisfaction, which was measured with the system usability scale (SUS), the mean was 53.88, with a range of 10 to 100. Figure 3 shows there was not evidence to support a statistically significant difference in SUS scores, $F(3, 57)=2.25$, $MSE=462.80$, $p=.092$, $\eta_p^2=.11$. None of the systems received SUS scores that were acceptable, as specified by Bangor, Kortum and Miller (2009).

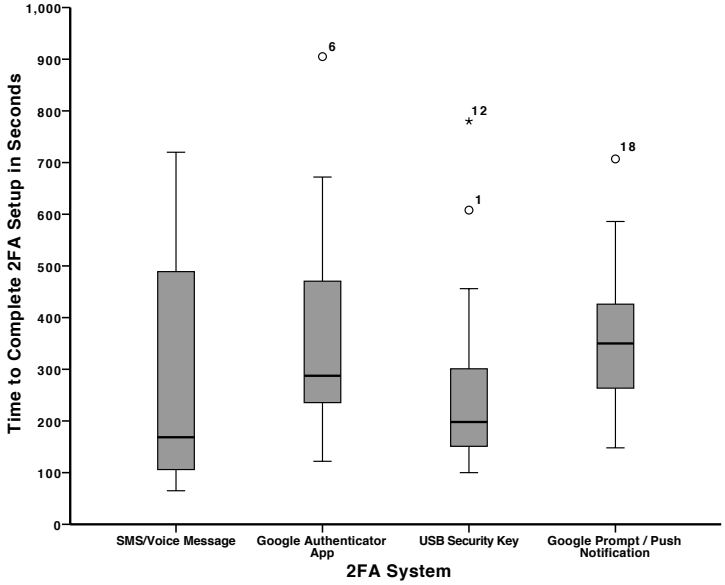

Figure 1. Time (seconds) to complete system setup as a function of 2FA method

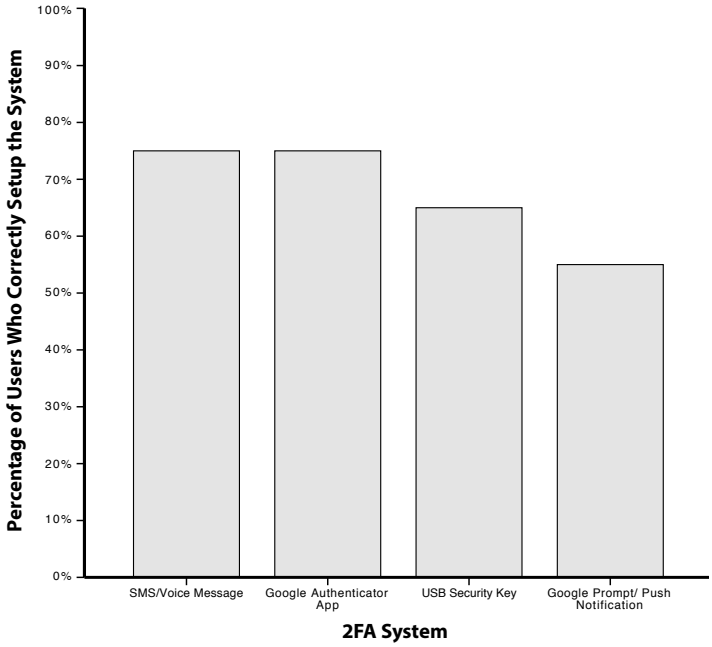

Figure 2. Percentage of users who correctly setup the system as a function of 2FA system

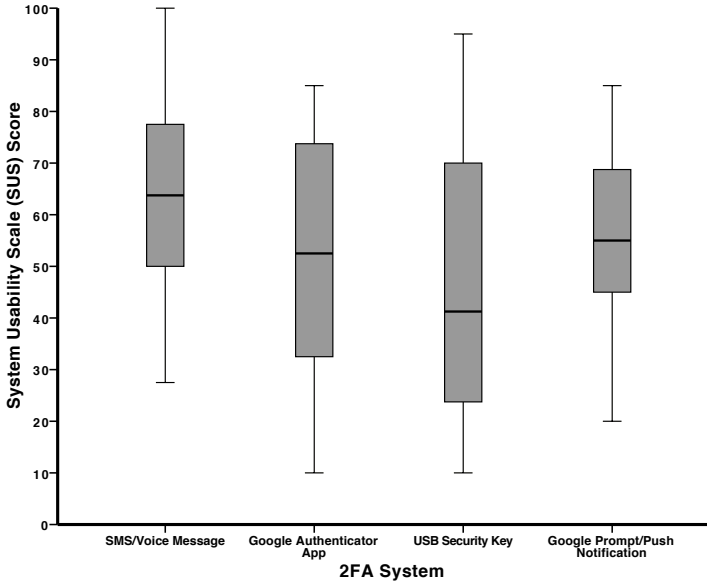

Figure 3. System Usability Scale (SUS) score as a function of 2FA system

2FA Login

Efficiency. It took participants a mean of 58 seconds, with a range of 14 to 590s, to login into their Gmail account using 2FA. As can be seen in Figure 4, there was not a reliable effect of login time, $F(1.62, 30.85) = 1.03$, $MSE = 8{,}074.39$, $p = .36$, $\eta_p^2 = .05$ (Greenhouse-Geisser corrected).

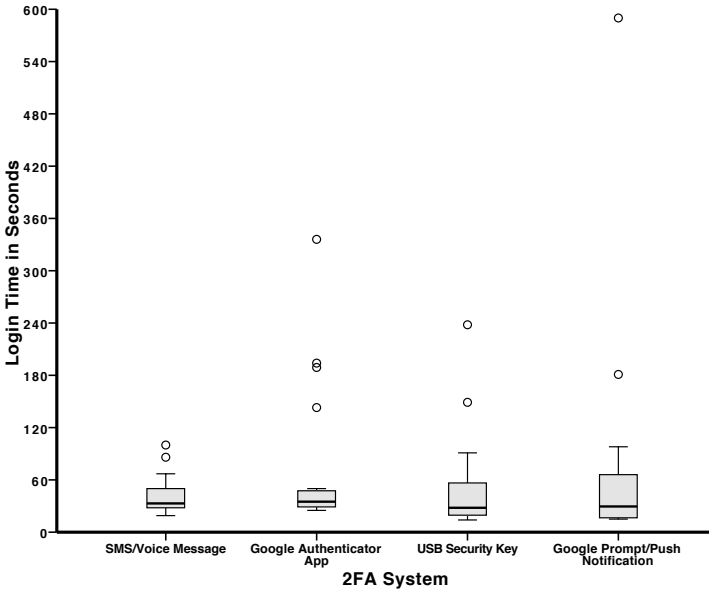

Figure 4. Total time to login (seconds) as a function of 2FA system

Effectiveness. On average, 95% of the participants were able to login to Gmail using 2FA methods, with a range of 85% to 100%. There was not a reliable effect, as illustrated in Figure 5, $F(1.54, 29.29) = 2.59$, $MSE = .08$, $p = .104$, $\eta_p^2 = .12$ (Greenhouse-Geisser corrected).

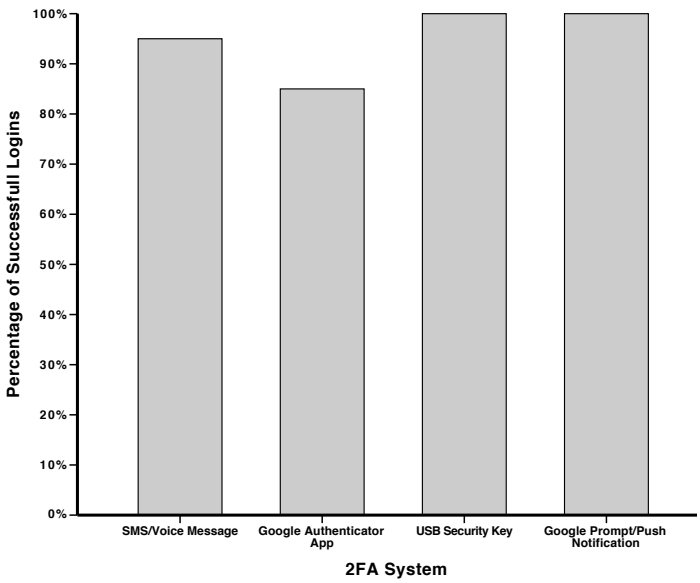

Figure 5. Percentage of successful logins as a function of 2FA method

Satisfaction. The mean SUS was 75.28, with a range of 10 to 100. There were some small, general differences across means, $F(3, 57) = 4.13$, $MSE = 309.50$, $p = .010$, $\eta_p^2 = .18$, meaning some systems were perceived to be more usable than others (see Figure 6). When direct comparisons were made, the security key was found to be more usable than the Google authenticator app, $t(19) = 2.56$, $p = .019$, Cohen's $d = 0.64$. In addition, Google Prompt (i.e., a push notification on the phone) was perceived to be more usable than the authenticator app, $t(19) = 3.01$, $p = .007$, Cohen's $d = 0.83$. Overall, the systems were deemed to have acceptable usability (Bangor, Kortum and Miller (2009).

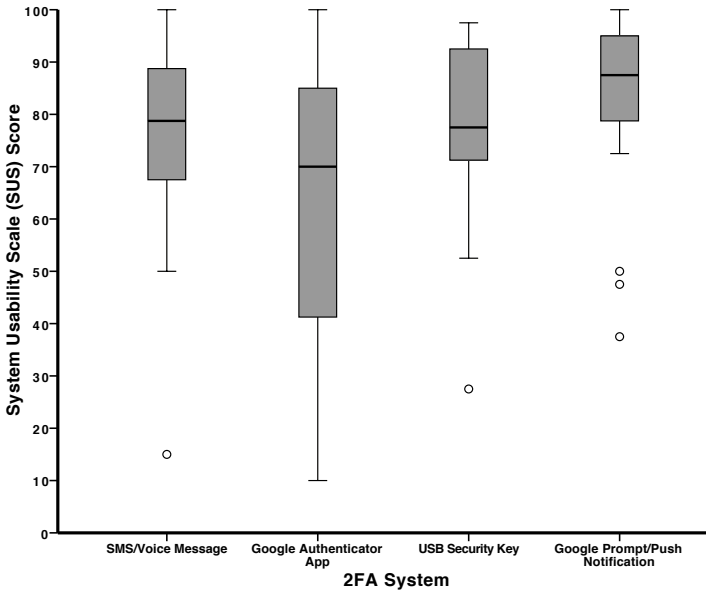

Figure 6. System Usability Scale (SUS) score as a function of 2FA method

When the three measures of efficiency, effectiveness, and satisfaction (SUS score) are compared across the setup and login tasks, it is revealed that it takes people a lot longer to setup 2FA ($\bar{x}$=314s, s=114.30) than to use it to login to the email system ($\bar{x}$=58s, s=58.51), $t(19)$=8.36, p<.001, 95% CI=191.86-320.06, Cohen's d=2.82. Fewer people were able to complete the setup task ($\bar{x}$=68%, s=0.24) than the login task ($\bar{x}$=95%, s=0.13), $t(19)$=5.40, p<.001, 95% CI=0.38-0.17, Cohen's d=1.40. Participants indicated that the login task was easier to complete (SUS $\bar{x}$=53.88, s=11.00) than the setup task (SUS $\bar{x}$=75.28, s=13.70), $t(19)$=6.06, p<.001, 95% CI=28.80-14.01, Cohen's d=1.72, which is consistent with the two objective measures. These findings are concerning, because if people are not able to setup the system, then they will not be able to benefit from the added security even if the login task is easier to complete.

DISCUSSION

The tested Google 2FA techniques are hard to use. In particular, the setup of 2FA was difficult—and in some cases impossible—to complete. This is especially problematic because even if a user can use 2FA once it is configured, an especially hard setup experience can prevent and/or discourage users from using the added security to keep their accounts safer in the first place. Beyond setup, everyday use of these 2FA systems can be tedious, with login times of over a minute hindering seamless usability, potentially resulting in user attrition over time.

There are several limitations to this study, some of which have already been acknowledged earlier in the paper—like the participants being undergraduates and mostly female. Future research can use participants that are more representative of all 2FA users and systematically study 2FA on different devices. Another issue is that the time-on-task and SUS measures were likely impacted by recoverable equipment failures. While equipment failures are never ideal when running a usability assessment, the equipment problems observed (e.g., a server not responding and fluctuating internet speed on some days) would still occur in the field and impact usability. These types of issues should be identified and investigated through future research so that they can be mitigated in the next generation of 2FA techniques.

In conclusion, even though 2FA has many security benefits, users may fail to take advantage of these if they view the set up and use of 2FA as overly onerous. Significant attention needs to focus on how to make these systems much easier to set up and use so that they will be embraced by the broad demographic that comprises Google users. The easier a system is to use, the more people will be willing to use it (Green, 2012). Finally, this study also reinforces the importance of a system being both secure *and* usable. If people cannot use a secure system, then the added security is not actually helping.

ACKNOWLEDGEMENTS

This work was supported in part by NSF grant CNS-1409401.

REFERENCES

Acemyan, C. Z., Kortum, P., Byrne, M. D., & Wallach, D. S. (2014). Usability of voter verifiable, end-to-end voting systems: Baseline data for Helios, Prêt à Voter, and Scantegrity II. *USENIX Journal of Election Technology and Systems (JETS)*, *2*(3), 26–56.

Acemyan, C. Z., Kortum, P., Byrne, M. D., & Wallach, D. S. (2015). From error to error: Why voters could not cast a ballot and verify their vote With Helios, Prêt à Voter, and Scantegrity II. *USENIX Journal of Election Technology and Systems (JETS)*, *3*(2), 1–25.

ArunPrakash, M., & Gokul, T. R. (2011). Network security-overcome password hacking through graphical password authentication. In *Proceedings of National Conference on Innovations in Emerging Technology, NCOIET'11* (pp. 43–48). IEEE.

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123.

Bevan, N. (2009). Usability. In L. LIU & M. T. ÖZSU (Eds.), *Encyclopedia of Database Systems* (pp. 3247–3251). Boston, MA: Springer US.

Brandom, R. (2017, July). Two-factor Authentication Is a Mess: It was supposed to be a one-stop security fix. What happened? *The Verge*. Retrieved from https://www.theverge.com/2017/7/10/15946642/two-factor-authentication-online-security-mess

Brooke, J. (1986). Usability engineering in office product development. In *Proceedings of the Second Conference of the British Computer Society, Human Computer Interaction Specialist Group On People and Computers: Designing for usability* (pp. 249–259). Cambridge University Press.

Byrne, M., Greene, K., & Everett, S. (2007). Usability of voting systems: Baseline data for paper, punch cards, and lever machines. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 171–180). ACM.

De Cristofaro, E., Du, H., Freudiger, J., & Norcie, G. (2013). A comparative usability study of two-factor authentication. *arXiv preprint arXiv:1309.5344*.

Green, B. D. (2012). Six Rules For Creating Products People Love. In *Six Rules For Creating Products People Love* (pp. 73–92). Bloomington: Author House.

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

International Organization for Standardization. (1997). *ISO 9241-16: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)-Part 16: Direct-manipulation Dialogues*.

Kortum, P. (2016). *Usability assessment: How to measure the usability of products, services, and systems*. Human Factors and Ergonomics Society.

Morley, K. (2016, February 10). Use the same password for everything? You're fueling a surge in current account fraud. *The Telegraph*.

Ong, T. (2018, January). Reddit now offers two-factor authentication to all. *The Verge*. Retrieved from https://www.theverge.com/2018/1/25/16931572/reddit-two-factor-authentication

Weir, C. S., Douglas, G., Carruthers, M., & Jack, M. (2009). User perceptions of security, convenience and usability for ebanking authentication tokens. *Computers & Security*, *28*(1-2), 47-62.

Zetter, K. (2016, April). The Critical Hole at the Heart of Our Cell Phone Networks. *Wired*. Retrieved from https://www.wired.com/2016/04/the-critical-hole-at-the-heart-of-cell-phone-infrastructure/