Behavior Ever Follows Intention?

A Validation of the Security Behavior Intentions Scale (SeBIS)

Serge Egelman[1], **Marian Harbach**[1], **Eyal Peer**[2]
[1]International Computer Science Institute, Berkeley, CA
[2]Bar-Ilan University, Ramat Gan, Israel
{egelman,mharbach}@icsi.berkeley.edu, eyal.peer@biu.ac.il

ABSTRACT

The Security Behavior Intentions Scale (SeBIS) measures the computer security attitudes of end-users. Because intentions are a prerequisite for planned behavior, the scale could therefore be useful for predicting users' computer security behaviors. We performed three experiments to identify correlations between each of SeBIS's four sub-scales and relevant computer security behaviors. We found that testing high on the *awareness* sub-scale correlated with correctly identifying a phishing website; testing high on the *passwords* sub-scale correlated with creating passwords that could not be quickly cracked; testing high on the *updating* sub-scale correlated with applying software updates; and testing high on the *securement* sub-scale correlated with smartphone lock screen usage (e.g., PINs). Our results indicate that SeBIS predicts certain computer security behaviors and that it is a reliable and valid tool that should be used in future research.

Author Keywords

Security behavior; measurement; user studies

ACM Classification Keywords

J.4. Social and Behavioral Sciences: Psychology; K.6.5. Management of Computing and Information Systems: Security and Protection

INTRODUCTION AND RELATED WORK

Understanding users' computer security attitudes is important for researchers because it helps to contextualize users' observed behaviors, as well as to potentially predict their future behaviors. For example, an understanding of a given user's attitudes may shed light on the likelihood that she will comply with a given security mitigation; a user with poor security attitudes is unlikely to take steps to comply. Similarly, if a user has strongly stated security intentions, yet exhibits poor security behaviors, this may indicate usability problems (i.e., Norman's "Gulf of Execution" [9]). Thus, when studying users' computer security behaviors, it is incumbent on researchers to evaluate the gap between attitudes and behaviors.

Within the privacy literature, researchers have long studied users' privacy attitudes: various privacy concerns scales have been developed to measure users' privacy preferences (e.g., [11, 6, 3, 7]). Measuring users' attitudes has allowed for the study of the "privacy paradox," which is the gap between self-reported attitudes and observed privacy-preserving behaviors [1]. For instance, Egelman *et al.* showed that more usable user interfaces can help close this gap [5].

One challenge to applying these methods to computer security behaviors is that up until very recently, there has not existed a reliable metric to gauge users' computer security attitudes. This changed in 2015, when Egelman and Peer introduced the Security Behavior Intentions Scale (SeBIS) [4], which is a 16-item instrument scored on a 5-point Likert scale that measures a computer user's self-reported intent to comply with "good" security practices spanning four dimensions:

- **Awareness**: Do users pay attention to contextual cues, such as the web browser URL bar or various security iconography?
- **Passwords**: Do users create unique passwords that exceed minimum requirements and are therefore difficult to crack?
- **Updating**: Do users apply software security updates in a timely manner?
- **Securement**: Do users secure their devices with secret codes, such as using smartphone secure lock screens (i.e., requiring a PIN) or password-protected screen savers on desktops and laptops?

Understanding computer security intentions is important, because intentions are a prerequisite of planned behavior [2]. In their original paper, Egelman and Peer showed how computer security intentions—as measured by SeBIS—are correlated with various well-studied psychological constructs, and that SeBIS exhibits high internal reliability [4]. However, one shortcoming of their work was that they did not demonstrate criterion validity; that is, does SeBIS actually predict computer security behavior? And if so, to what extent?

We are unaware of any prior published research that has attempted to correlate the four SeBIS sub-scales (or any other measure of computer security intentions) with observed computer security behaviors. Thus, we contribute the following:

- We show that testing high on the *awareness* sub-scale is significantly correlated with the ability to correctly identify a phishing website.

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05...$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858265

- We show that testing high on the *passwords* sub-scale is significantly correlated with the ability to create passwords that cannot be cracked in a reasonable amount of time.
- We show that testing high on the *updating* sub-scale is significantly correlated with applying important software updates in a timely manner.
- We show that testing high on the *securement* sub-scale is significantly correlated with the use of smartphone locking mechanisms (e.g., PINs or drawing a pattern).

In performing this study, we conducted three experiments. Our first experiment used a single cohort of online participants and was designed to examine the *awareness* and *passwords* sub-scales. Our second experiment used a new cohort of online participants to examine the *updating* sub-scale, and our third experiment was a field study of smartphone users in order to examine the *securement* sub-scale. Examining each of SeBIS's sub-scales involved measuring very different types of behaviors, and therefore one single experiment is unlikely to examine all four dimensions with any sort of ecological validity. As a result, the remainder of this paper is divided into three sections based on these experiments, concluding with discussion.

AWARENESS AND PASSWORDS

Our first experiment was performed entirely online using Amazon's Mechanical Turk and was designed to examine the *awareness* and *passwords* SeBIS sub-scales.

The SeBIS *awareness* dimension is designed to measure the extent to which users claim to be aware of various contextual cues. This includes indicators such as a web browser's URL bar, which may yield information about the website being visited, such as whether it uses encryption or whether the domain name matches the user's expectations. The latter can be used to detect phishing websites. As such, we designed an experiment to test the following hypothesis:

> $H_{awareness}$: *Users who correctly use the URL bar to identify a phishing website will test significantly higher on the awareness sub-scale than users who do not correctly identify a phishing website.*

The SeBIS *passwords* dimension is designed to measure the extent to which users claim to create strong and unique passwords. We examined the following hypothesis:

> $H_{passwords}$: *Users who create passwords that cannot easily be cracked will test significantly higher on the passwords scale than users whose passwords can be cracked.*

Methodology

We recruited participants from Amazon's Mechanical Turk to complete an online task in two parts, with a two-week gap in between. In this section, we describe how we used this experiment to examine behaviors that we hypothesized would correlate with the SeBIS's *awareness* and *passwords* sub-scales. We also describe the followup task, which included administering SeBIS and collecting demographics.

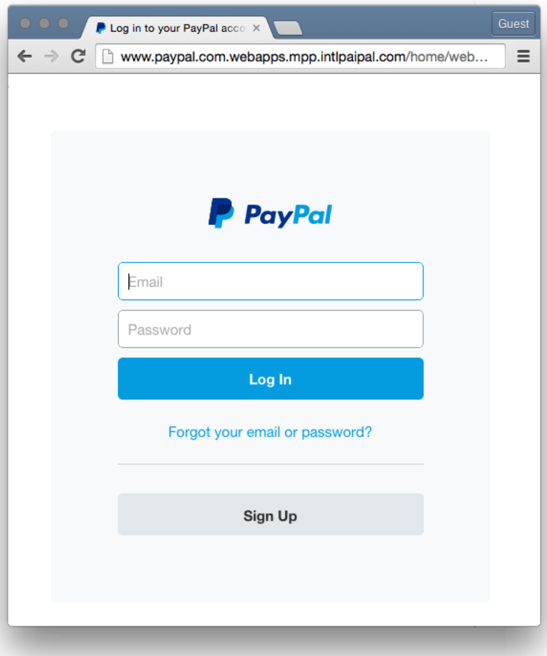
Log in to your PayPal acco
Guest
www.paypal.com.webapps.mpp.intlpaipal.com/home/web...
PayPal
Email
Password
Log In
Forgot your email or password?
Sign Up

Figure 1. We showed participants a screenshot of this website and asked them to describe it in an open-ended response. Those who used the included URL bar to realize that it was a phishing website scored significantly higher on the SeBIS *awareness* sub-scale.

Awareness

As part of a different study, participants completed a series of psychometric scales in exchange for $1.00. To prevent fatigue and detect attempts at cheating, between each scale, we displayed a screenshot of a website and asked them to describe the website pictured using the text box provided. They performed this task three times; the first two screenshots were of the legitimate Amazon and Twitter websites, whereas the third screenshot was of a phishing website spoofing PayPal (Figure 1). The only way of identifying the phishing website was by examining the URL bar in the screenshot.

Passwords

At the end of this set of tasks, we asked each participant to create a password, which we told them they would need to enter when returning two weeks later to complete the second half of the study. Because we wanted to examine the extent to which participants' passwords would go beyond any minimum requirements, we set the requirements to be relatively weak: 6 characters with at least 1 digit or symbol. (We also wanted to see whether participants would reuse existing passwords, and therefore chose minimum requirements that were likely to be weaker than those found on other websites.)

The reason for the two-week gap was that we were concerned that if we immediately presented participants with SeBIS right after the website identification and password creation tasks, the scale might be biased by participants' recollections of their behaviors. Similarly, we were also concerned that participants may discuss the security nature of the study in various online forums, thereby tainting future participants.

As a result, two weeks after participants had completed the initial task, we used Amazon's internal messaging system to invite them to complete SeBIS for a $2.00 bonus payment.

To analyze the passwords participants created in the first part, we uploaded them to CMU's Password Guessability Service (PGS)[1] [12], which hashed the passwords and then attempted to crack them using four different approaches. We used the minimum number of guesses it took to crack each password across all four approaches as our strength metric. Ur *et al.* used this approach and concluded that it represents "a conservative estimate for the performance of an experienced professional attacker who uses all techniques, wordlists, tools, and dynamic updates at his/her disposal" [12]. Additionally, in our exit survey (administered at the same time as SeBIS), we asked participants whether they had used this password on other websites.

Exit Survey and Demographics

Two weeks after completing the first part of the study, 718 participants returned (we discarded the data of 260 participants who did not attempt to return). We gave them three attempts to enter their passwords, after which the survey simply let them in. We decided on this policy so that we could balance the conflicting requirements of testing password memorability with minimizing attrition from the study as a whole due to forgotten passwords. That is, 206 participants (28.7% of 718) could not remember their passwords two weeks later within three attempts, and therefore we removed their data from the *password* sub-scale evaluation—but still used it as part of the *awareness* evaluation—because we could not know the reasons for their authentication failures (e.g., they may not have created passwords that they intended to remember, and therefore these passwords would not be representative of passwords used in real life), whereas failure to remember the password was unlikely to impact their non-password behaviors. Thus, our password analysis included 512 subjects who successfully recalled their passwords.

We first examined the internal validity of participants' SeBIS responses and found that they were consistent with the original paper [4]: Cronbach's $\alpha_{total} = 0.81$, $\alpha_{securement} = 0.77$, $\alpha_{passwords} = 0.74$, $\alpha_{awareness} = 0.65$, $\alpha_{updating} = 0.70$. Regarding demographics, our 718 participants were split roughly evenly between 370 males (51.5%) and 348 females (48.5%), with a mean age of 35 (range of 19-69). All participants were from the U.S. and had completed at least 500 previous tasks on Mechanical Turk with an approval rating of 95% or higher [10]. The demographics of our subset of 512 participants who successfully remembered their passwords did not significantly differ from the superset.

Results

Awareness

To examine *awareness* behaviors, two independent coders examined the open-ended responses to the phishing website description question for correctness (Cohen's $\kappa = .83$) and identified 22 participants (3.1% of 718) who correctly identified the phishing website. While this number was small, these 22 participants had mean *awareness* sub-scale scores that were larger than those who were unable to identify the phishing website: 4.31 vs. 3.68. A t-test (data was normally distributed) indicated that this difference was statistically significant ($t = 5.22$, $p < 0.0005$) with a large effect size (Cohen's $d = 0.92$). Upon correcting for multiple testing, none of the other sub-scales yielded statistically significant differences. Thus, $H_{awareness}$ is supported.

Passwords

Of the 512 passwords that we uploaded to PGS for cracking, 75 could not be cracked (14.7%). Participants whose passwords could be cracked had average *passwords* sub-scale scores of 3.21, whereas participants whose passwords could not be cracked had average sub-scale scores of 3.56. This difference was statistically significant ($t = 3.47$, $p < 0.001$) with a medium effect size (Cohen's $d = 0.41$). This effect was not present with any of the other SeBIS sub-scales.

In the exit survey, 196 participants (38.3% of 512) reported that they reused an existing password that they use on other websites. Not surprisingly, these participants scored significantly lower on the *passwords* sub-scale than participants who claimed to have created a new password: 2.94 vs. 3.46 ($t = 6.94$, $p < 0.0005$), respectively. This effect was relatively large (Cohen's $d = 0.63$). However, we were surprised to find that participants who claimed to create new passwords scored significantly higher on each of the three other sub-scales, with medium effect sizes, even after correcting for multiple testing:

- Securement: $t = 3.66$, $p < 0.0005$, $d = 0.32$
- Awareness: $t = 4.79$, $p < 0.0005$, $d = 0.43$
- Updating: $t = 3.87$, $p < 0.0005$, $d = 0.35$

These results suggest that not only is $H_{passwords}$ supported, but that users who claim to create unique passwords intend to engage in better security practices across the other areas measured by SeBIS.

UPDATING

The SeBIS *updating* dimension is designed to measure the extent to which users claim to apply security updates in a timely manner. We examined the following hypothesis:

> $H_{updating}$: *Users who promptly apply software updates will test significantly higher on the updating scale than users who delay applying software updates.*

During our previous experiment, we wanted to examine the *updating* sub-scale after we serendipitously discovered that critical operating system updates—those requiring a reboot—were released between the first and second parts of our experiment. However, only 50 subjects were eligible for these updates, which did not yield enough statistical power. As a result, we decided to perform an additional experiment to specifically examine updating behaviors.

Methodology

On October 20, 2015, three weeks after Mac OS X 10.11 came out (requiring manual installation), we deployed a new survey on Mechanical Turk, which was only available to Mac

[1] https://pgs.ece.cmu.edu/

users who had previously not completed SeBIS. We asked participants to complete SeBIS as well as to provide demographics and the model and owner of the Mac they were using. We screened for Mac OS based on the `user-agent` string. We paid participants $0.50, and they took about 3 minutes on average. We excluded participants who were not completing the survey on their own Mac (self-report) and who were ineligible for the update, based on the model information (as copied from the "About This Mac" dialogue). Finally, we determined whether or not the update was installed based on the version of Mac OS X reported by their web browser's `user-agent` string.

Result

Of the 359 participants who successfully completed our survey, 281 (56.2% female, median age = 28) were eligible for the update. We found a statistically significant difference between participants who had installed the update within three weeks (24.2% of 281) vs. those who had not ($M = 3.52$ vs. 3.02, $t = 4.11$, $p < 0.0001$), with a medium effect size (Cohen's $d = 0.59$). Thus, $H_{updating}$ is supported.

SECUREMENT

The SeBIS *securement* dimension measures the extent to which users claim to secure their electronic devices when not in use. This amounts to enabling password-protected screensavers on desktops and laptops, manually locking screens when stepping away, enabling secure lock screens on mobile devices (e.g., requiring a PIN, pattern, or biometric), and so forth. We examined the following hypothesis:

> $H_{securement}$: *Users who employ secure lock screens on smartphones will test significantly higher on the securement scale than users not employing secure lock screens.*

Methodology

We performed a third experiment using an entirely different sample recruited from PhoneLab [8].[2] PhoneLab is an experimental platform maintained by researchers at the University of Buffalo, consisting of a panel of about 200 Android smartphone users who have agreed to have their smartphones instrumented for research purposes. We deployed instrumentation to the PhoneLab panelists that collected data about whether or not they employed a secure lock screen on their Android device. We recorded whether each participant unlocked his or her phone via either entering a PIN or drawing a pattern, or using the insecure "slide-to-unlock" method, which does not require entering a secret (i.e., typing a PIN or drawing a pattern). This instrumentation occurred transparently to participants and therefore was unlikely to prime them to think about security.

After collecting our data on participants' unlock methods, we emailed them a link to an online exit survey. The exit survey consisted of SeBIS and various demographic questions. As an incentive to complete the survey, we created a drawing for a $100 Amazon gift card. Of the 71 participants who completed the exit survey, 50.1% were female and the median age was 35, with a range of 19-70 years.

Result

Our 71 participants were split between PIN users (25.4% of 71), pattern users (25.4% of 71), and those using the insecure slide-to-unlock method (49.3% of 71). While we observed no difference in SeBIS *securement* scores between those using either PINs or patterns, we did observe that the mean SeBIS scores of slide-to-unlock users was a full point below users of either of the two secure unlock methods: 2.75 vs. 4.25. A Wilcoxon rank sum test (SeBIS scores were not normally distributed, which we attribute to the relatively small sample size) indicated that this difference was statistically significant ($W = 169$, $p < 0.0005$) with a large effect size ($r = 0.63$). Thus, $H_{securement}$ is supported.

DISCUSSION

In this study, we examined the criterion validity of the Security Behavior Intentions Scale (SeBIS) to see whether participants' self-reported security intentions, as measured by the scale, predicted their security behaviors. We observed that participants' SeBIS sub-scale scores for all four sub-scales were predictive of very specific security behaviors.

Our results show that SeBIS does in fact predict particular security behaviors, and in many cases, the effect sizes were large. At the same time, our study had several limitations. The behaviors that we examined were relatively narrow in scope and were also chosen to align with concepts specifically mentioned in SeBIS (e.g., SeBIS directly asks about the locking behavior that we observed). Further research is therefore needed to examine a wider spectrum of security-related behaviors. For instance, are individuals who test high on *securement* more likely to log out of websites? Do scores on the *passwords* scale predict usage of password managers? Are individuals testing low on *awareness* more likely to delegate computer maintenance to others? Are individuals testing high on *updating* more discriminating with regard to provenance when choosing software to install?

Furthermore, we only examined password creation behaviors for relatively low-risk accounts; it is therefore not clear how these behaviors might generalize to high-risk accounts. Participants' password reuse was also self-reported; future studies may be needed to further examine these behaviors.

Despite the shortcomings outlined above, we contribute to the literature by demonstrating that SeBIS is predictive of certain computer security behaviors. While further studies are needed, we conclude that SeBIS is a useful tool for researchers who need to assess users' computer security attitudes and related behaviors.

ACKNOWLEDGEMENTS

We would like to thank the PhoneLab team at the University of Buffalo and the Password Guessability Service at Carnegie Mellon University for their support in gathering the data. This work was supported by the FITweltweit program of the German Academic Exchange Service (DAAD); U.S. National Science Foundation awards CNS-1343433, CNS-1343451, and CNS-1528070; U.S. National Security Agency Lablet contract H98230-14-C-0140; and U.S.-Israel Binational Science Foundation award 2014626.

[2] https://www.phone-lab.org/

REFERENCES

1. A. Acquisti and R. Gross. 2006. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. In *Privacy Enhancing Technologies Workshop (PET '06) (Lecture Notes in Computer Science)*, Vol. 4258. Springer-Verlag, Berlin / Heidelberg, Germany, 36–58.

2. Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes* 50, 2 (1991), 179–211.

3. Tom Buchanan, Carina Paine, Adam N Joinson, and Ulf-Dietrich Reips. 2007. Development of measures of online privacy concern and protection for use on the Internet. *Journal of the American Society for Information Science and Technology* 58, 2 (2007), 157–165.

4. Serge Egelman and Eyal Peer. 2015. Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SeBIS). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2873–2882. DOI: http://dx.doi.org/10.1145/2702123.2702249

5. Serge Egelman, Janice Tsai, Lorrie Faith Cranor, and Alessandro Acquisti. 2009. Timing is Everything?: The Effects of Timing and Placement of Online Privacy Indicators. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 319–328. DOI: http://dx.doi.org/10.1145/1518701.1518752

6. Ponnurangam Kumaraguru and Lorrie Faith Cranor. December, 2005. *Privacy Indexes: A Survey of Westin's Studies*. Technical Report CMU-ISRI-5-138. Carnegie Mellon University. http://reports-archive.adm.cs.cmu.edu/anon/isri2005/abstracts/05-138.html.

7. Naresh K. Malhotra, Sung S. Kim, and James Agarwal. 2004. Internet Users' Information Privacy Concerns (IUIPC): The Construct, The Scale, and A Causal Model. *Information Systems Research* 15, 4 (December 2004), 336–355.

8. Anandatirtha Nandugudi, Anudipa Maiti, Taeyeon Ki, Fatih Bulut, Murat Demirbas, Tevfik Kosar, Chunming Qiao, Steven Y. Ko, and Geoffrey Challen. 2013. PhoneLab: A Large Programmable Smartphone Testbed. In *Proceedings of First International Workshop on Sensing and Big Data Mining (SENSEMINE'13)*. ACM, New York, NY, USA, Article 4, 6 pages. DOI: http://dx.doi.org/10.1145/2536714.2536718

9. Donald A. Norman. 1986. Cognitive Engineering. In *User Centered System Design: New Perspectives on Human-Computer Interaction*, Donald A. Norman and Stephen W. Draper (Eds.). Lawrence Erlbaum Associates, London, Chapter 3, 31–62.

10. Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods* 46, 4 (December 2014), 1023–1031.

11. Sören Preibusch. 2013. Guide to Measuring Privacy Concern: Review of Survey and Observational Instruments. *International Journal of Human-Computer Studies* 71, 12 (Dec. 2013), 1133–1143. DOI: http://dx.doi.org/10.1016/j.ijhcs.2013.09.002

12. Blase Ur, Sean M. Segreti, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Saranga Komanduri, Darya Kurilova, Michelle L. Mazurek, William Melicher, and Richard Shay. 2015. Measuring Real-World Accuracies and Biases in Modeling Password Guessability. In *24th USENIX Security Symposium (USENIX Security 15)*. USENIX Association, Washington, D.C., 463–481. https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/ur