Toward Accurate Prediction of Security Behavior via Comprehensive Scales

Yukiko Sawaya
yu-sawaya@kddi-research.jp
KDDI Research, Inc.

Takamasa Isohara
ta-isohara@kddi-research.jp
KDDI Research, Inc.

Mahmood Sharif
mahmoods@cs.tau.ac.il
Tel Aviv University

Abstract

Psychometric security and privacy scales can enable various crucial tasks (e.g., measuring changes in user behavior over time), but, unfortunately, they often fail to accurately predict actual user behavior. We hypothesize that one can enhance prediction accuracy via constructing more comprehensive scales measuring a wider range of factors related to security and privacy. This article describes our preliminary efforts toward validating this hypothesis by developing a more comprehensive security scale measuring end-user behavior intentions. More precisely, we explain how we formed an initial set of items to include in the scale, and present a follow-up online user study (*n*=299) to refine the scale and uncover its latent structure (i.e., characterize the sub-scales that form it). Our work led to the development of a scale with ∼44% more items and 25% more factors than SeBIS, a widely accepted security behavior intentions scale. We close the article by discussing our plans to finalize the scale and test our hypothesis.

1 Introduction

Billions of people worldwide regularly spend a significant amount of their time online or interacting with technological devices. In fact, a recent report shows that the average Internet user spend more than six hours per day online [11]. During this time, users constantly face decisions that directly impact their security and privacy, ranging from configuring permissions to allow or prevent newly installed apps from accessing certain information to selecting options to control who can view their activities on social media.

In return, researchers have attempted to develop a rigorous understanding of users' privacy attitudes, behaviors, concerns, and preferences to inform the design of systems that serve the users best. Among others, psychometric scales such as the Privacy Concerns Scale (PCS) [3], Internet Users' Information Privacy Concerns (IUIPC) [13], the Westin Index [12], and the Security Behavior Intentions Scale (SeBIS) [8] have been proposed as affordable and scalable means to learn about users' security and privacy attitudes, concerns, and bahaviors. Conceptually, these scales can be useful for various goals, including enabling us to configure systems to safe defaults respecting users' preferences (e.g., configuring sharing and tagging policies on social networks [23]); bootstrapping personalized defenses to usable states (e.g., ones to automatically enable or block tracking per users' preferences [15]); raising users' privacy awareness (e.g., when they underestimate certain risks [10]); or measuring changes in users' behaviors, concerns, or preferences over time (e.g., due to interventions such as user education or the implementation of new security and privacy features [6, 9]).

However, unfortunately, *prior scales are often found to be poor predictors of actual behavior*. For example, Woodruff et al. found no correlation between the Westin Index and respondents' privacy behavior in certain scenarios, such as ones probing whether they would be willing to sell their medical records for a certain fee [24]. Similarly, our work has found that two of three IUIPC sub-scales do not explain users' likelihood to share their private data, while the third had markedly weaker explanatory power than other factors (e.g., the party with whom the data is shared) [22].

A seeming counterexample is the work of Egelman et al. who showed that scores on SeBIS—a 16-item scale composed of four sub-scales measuring dimensions related to proactive awareness, password generation, updating, and device securement—are correlated with users' security-related behavior [7]. More precisely, Egelman et al. have shown that study participants who scored highly on proactive awareness were less likely to be deceived by phishing; participants who attained high scores on the password-generation sub-scale created harder-to-guess passwords; participants who achieved high scores on the updating sub-scale were more likely to update their operating system within a short period of a new version's releaset; and those who scored more highly on device securement were more likely to lock their phones using PINs or patterns. Nonetheless, prior work has also found that while SeBIS responses could predict users' exposure to malicious websites, they were significantly less accurate than behavioral features at doing so (∼20% lower area under the receiver operating characteristic curve) [21]. Thus, it remains unclear if users' responses to standard scales such as SeBIS can predict security behavior as accurately as other behavioral indicators (e.g., one learned via telemetry), and to what extent, if any, one can improve the prediction accuracy of such scales.

We hypothesize that the omission of critical factors that may directly impact users' security and privacy behavior may harm scales' ability to predict actual behavior. Said differently, we expect that *scales that cover a more comprehensive set of factors related to security and privacy can predict users' behavior more accurately*. This article reports on our efforts developing such comprehensive scale for measuring security behavior toward putting this hypothesis to the test. Specifically, we present how we created an initial set of items to include in the scale (Sec. 2). We then report on the process we followed to refine the scale and characterize its latent construct, and compare the resulting scale with SeBIS (Sec. 3). We conclude by discussing our preliminary results and describing our future work, including how we plan to corroborate our hypothesis (Sec. 4).

2 Initial Scale's Items

Similarly to SeBIS [8], we based our scale's questions on widely recommended security advice to measure end-users' compliance intentions. Yet, unlike SeBIS, which started from 30 advised behaviors, we used a richer, more exhaustive pool. Specifically, we built off of Redmiles et al.'s work [18] to form an initial comprehensive set of advice. Their work analyzed >2,000 documents containing security and privacy advice, and identified 374 advice items that are often suggested to users. These items pertain to twelve categories, ranging from account security to network security, and from password creation and management to anti-viruses. The categories are mostly mutually exclusive, except for four advice items that appear in two categories each. Redmiles et al. surveyed 41 security and privacy experts to assess the advice items' accuracy (i.e., whether following them is conducive to security and privacy) and perceived utility (i.e., the expected risk reduction due to compliance with the advice). Furthermore, they asked experts to prioritize items, only to find out that there is no widely acceptable prioritization among experts—more than 50% of advice appeared at least once in experts' top-ten most recommended advice items.

While comprehensive, basing the scale's items on all 374 advice items would result in a long, impractical questionnaire: it would be prohibitive to complete the questionnaire in a reasonable amount of time, and participants are likely to become less engaged and drop out in the middle, thus harming data quality [2]. Therefore, we sought to select a subset of advice to include in our questionnaire. Particularly, we applied the following criteria for advice selection:

1. **High accuracy**: We ensured that all experts surveyed agreed that the advice items are accurate.
2. **High utility**: The advice selected had high perceived risk reduction. Specifically, we picked items whose perceived risk reduction was $\geq$40%—the median perceived reduction among all advice.
3. **High priority**: We excluded advice items that were not assigned high priority, only keeping items in the 50th percentile. Applying criteria 1–3 resulted in a marked decrease in the number of advice items, with 103 items surviving the selection process.
4. **Security relevance and wide applicability**: Two researchers manually and independently examined the remaining advice items and the documents they have originally appeared in to identify ones that are *1)* related to personal security and privacy (e.g., excluding items concerning giving advice to others); and *2)* applicable to a wide variety of users (e.g., not only ones who have kids or employed by technological enterprises). Overall, they found 30 items that do not satisfy the criteria, narrowing down the list of advice items to 83. The coders had substantial inter-coder agreement (Cohen κ=0.69) [14], and resolved disagreements manually by meeting to discuss and agree on definitions.
5. **Non-redundant**: As a final step, we identified and removed redundant advice. Over two consecutive meetings, two researchers clustered the advice items together according to their similarity, and picked a representative item for each cluster.

After applying all criteria, we remained with 45 initial advice items to include in our scale. Interestingly, these items covered 15 of the 16 advice items originally included in SeBIS. As a final step, in line with SeBIS, we rephrased the advice items to statements assessing frequency at which respondents follow recommended advice. Accordingly, responses to the scale are reported on a 5-point Likert scale: *Never (1), Rarely (2), Sometimes (3), Often (4), and Always (5).*

3 Refinement and Factor Analysis

3.1 Study Design

Following the recommended steps for scale development of Carpenter [4] and Netemeyer et al. [17], and following the development procedure of SeBIS, we designed an online user study to collect responses on our initial scale, refine it, and explore its latent construct via exploratory factor analysis (EFA). We asked participants in our study to fill out a survey composed of three primary parts. The first part assessed the participants' propensity to follow recommended advice. Here, participants reported the frequency at which they followed the

45 initial advice items described in Sec. 2. Besides answering questions on a Likert scale, we gave participants an "N/A" response option to find whether certain items are not widely applicable to our target audience. We were concerned that participants' responses were influenced by their desire to appear more socially acceptable [5]. Therefore, in the second part, we measured social desirability to test whether participants' willingness to appear more socially acceptable correlated with their answers to our survey questions. Particularly, to measure social desirability, we asked participants to complete a conventional 13-item version of the Marlowe-Crowne social desirability scale [5, 19, 19]. Finally, we asked participants basic demographic questions. Our study was reviewed and approved by our institution's ethics board.

We took several measures to maintain internal and external validity, and ensure data quality. To mitigate ordering effects, we presented the questions of the first and seconds parts in a randomized order. Additionally, to avoid selection bias, we advertised our study as one exploring technology perceptions, similarly to Abrokwa et al. [1]. Finally, we added attention questions at the beginning of the study, notified participants that failed them once, and precluded participants that failed them twice from completing the study.

3.2 Participants

We recruited participants via Prolific,[1] an online crowdsourcing platform. We opened our study to participants from the United States who are at least 18 years old. Additionally, we used Prolific's internal functionality to collect data from a population-representative sample, resulting in a sample whose ethnicity, gender, and age distribution reflects the general United States population's. A total of 307 participants started the study, and eight dropped out due to failing the attention question. Thus, overall, 299 participants completed our study, leading to >5:1 response-to-item ratio, as recommended [4]. The average participant's age was 45.6 (±16.0) years and 47.8% of the participants reported themselves as males. It took participants an average of 9.3 minutes to complete the survey, and they were compensated 1.6 GBP (≈2.0 USD) for participating.

3.3 Data Analysis and Result

We followed standard processes [4] to analyze the collected data. First, we examined how often questions received "N/A" responses. No question stood out as widely non-applicable, thus we decided to map "N/A" responses to "Never" (i.e., 1). Repeating the analysis using imputations to replace missing values [16], instead of fixed mapping, has led to consistent findings.

Secondly, we identified and removed items that exhibited ceiling (μ >4.0) or floor (μ <2.0) effects, or low variance (σ <1.0), as they have little utility in a scale. Second, we removed items that do not exhibit high total-item correlation ($\geq$0.3), as they do not measure the same construct as other items. This process resulted in the removal of 19 items. None of the items, including those removed, were correlated with the social desirability measure, indicating that participants' responses were not influenced by a social desirability bias.

Thirdly, we verified the factorability of the data by running Bartlett's test of sphericity, the Kaiser-Meyer-Olkin test of sampling adequacy, and inspecting the inter-item correlation matrix. We found that all statistics lie within the recommended ranges [4].

Finally, we performed EFA, using *principal component analysis* (PCA) and the Varimax rotation method [4], similarly to SeBIS [7]. Using standard procedures (including parallel analysis and optimal coordinates) [4], we set the number of latent factors as five—i.e., the scale we developed contains five sub-scales measuring five dimensions. The five factors we extracted explain over 53.6% of the variance.

None of the items exhibited low factor loadings (roughly, correlation) that warranted removal, however, we removed three items whose largest factor loading was not 20% higher than the second largest factor loading [4]. Hence, eventually, we remained with 23 items. We named the factors according to the themes of the items that belong to them. The factors and their corresponding items are reported in Table 1.

One can immediately see that our scale has more items (23 vs. 16) and more factors (5 vs. 4) than SeBIS. Interestingly, one of SeBIS' factors (password generation) is subsumed by a more general factor we have (account and data securement), while two of our factors (anti-virus and encryption) are not represented in SeBIS. In summary, we can conclude that we managed to develop a scale with higher coverage than SeBIS.

Perhaps surprisingly, we also found out that several SeBIS items (e.g., ones measuring how often users verify that information is sent of HTTPS, or whether they lock their devices) were excluded from our scale due to ceiling effects. We conjecture that this may be explained by changes in users' behavior in the seven years that have passed between SeBIS' publication [8] and our study. This conjecture, in fact, is marginally supported by our recent observations from running a variant SeBIS with a large sample (n ≈5,000) of users located in Japan, finding that participants reported significantly higher propensity of following certain security behavior, such as device locking, compared to six years ago [20].

4 Discussion and Future Work

In this article, we have described our efforts to develop a more comprehensive security scale than SeBIS. Nonetheless, although we have made significant progress constructing the scale, it is not yet finalized. Most notably, as a next step, we need to collect responses to the refined scale's questions and conduct a confirmatory factor analysis as well as reliability

[1] https://prolific.co

#	**Factor 1: Encryption (16.57% of variance explained; λ=3.81)**	μ	σ
1.1	I encrypt my email contents when sending sensitive information (e.g., banking and health information or social security number)	2.52	1.56
1.2	I validate the digital certificates on the websites I visit	2.64	1.38
1.3	I review the validity of my root certificates	2.12	1.40
1.4	I validate the digital signatures files before opening them	2.68	1.46
1.5	I encrypt my devices' disks to keep my data confidential	2.38	1.48
1.6	I safely store my private key for email encryption	2.60	1.68
#	**Factor 2: Proactive awareness (13.14% of variance explained; λ=3.02)**	μ	σ
2.1	I verify whom I communicate with online (via email or online messaging apps) is really the person I intend to	3.92	1.18
2.2	I verify links (e.g., in the URL bar or by mouseover) to ensure that I am accessing intended websites	3.83	1.19
2.3	I check the extensions (e.g., .exe, .pdf) of files I download	3.93	1.25
2.4	I turn on download notifications in my browsers	3.69	1.45
2.5	I report account breaches or losses to the appropriate people	3.45	1.57
2.6	I disable auto-run to prevent potentially malicious downloaded programs from running	3.61	1.55
#	**Factor 3: Account and data securement (9.92% of variance explained; λ=2.28)**	μ	σ
3.1	I avoid storing data that I do not need	3.71	1.12
3.2	I back up the data on my devices	3.68	1.18
3.3	I lock my computer when I am away from it	3.75	1.43
3.4	I select the strictest security settings (e.g., app permissions or browser options) that are practical	3.68	1.08
3.5	I select hard-to-guess passwords (with multiple character types, without dictionary words, etc.)	3.99	1.11
3.6	I select different passwords for different accounts and devices	3.93	1.09
#	**Factor 4: Anti-virus (8.95% of variance explained; λ=2.06)**	μ	σ
4.1	I scan attachments for viruses before downloading or opening them	3.51	1.43
4.2	I verify that my anti-virus software is up-to-date	3.77	1.30
4.3	I install anti-virus software when setting up my devices	3.91	1.31
#	**Factor 5: Updates (7.30% of variance explained; λ=1.68)**	μ	σ
5.1	I turn on automatic updates for devices and applications upon installation	3.56	1.22
5.2	When I am prompted about a device or software update, I immediately install it	3.30	1.13

Table 1: The final items included in our scale and the related factors (i.e., sub-scales) uncovered by EFA. For each item, we report the mean (μ) and standard deviation (σ), where responses were collected on a 5-points Likert scale (see Sec. 2). For each factor, we report the percentage of variance explained and associated eigenvalue from PCA.

analysis to confirm the latent construct we have identified and validate the scale's reliability [4, 17].

More crucially, we intend to use our scale to test the hypothesis put forward in Sec. 1, suggesting that more comprehensive security scales are more predictive of actual user behavior. To do so, we plan to recreate parts of Egelman et al.'s experiments [7] and assess which of SeBIS and our scale can predict behavior (e.g., identifying phishing or updating devices) more accurately, and by how much. We will also seek to compare SeBIS and our scale's ability to predict additional types of security behavior, such as exposure to malicious content, and contrast them with behavioral features.

References

[1] Desiree Abrokwa, Shruti Das, Omer Akgul, and Michelle L. Mazurek. Comparing security and privacy attitudes among US users of different smartphone and smart-speaker platforms. In *Proc. SOUPS*, 2021.

[2] Ioannis Andreadis and Evangelia Kartsounidou. The impact of splitting a long online questionnaire on data quality. In *Proc. Survey Research Methods*, 2020.

[3] Tom Buchanan, Carina Paine, Adam N Joinson, and Ulf-Dietrich Reips. 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):157–165, 2007.

[4] Serena Carpenter. Ten steps in scale development and reporting: A guide for researchers. *Communication Methods and Measures*, 12(1):25–44, 2018.

[5] Douglas Crowne and David Marlowe. A new scale of social desirability independent of psychopathology. *Journal of consulting psychology*, 24:349–54, 09 1960.

[6] Sauvik Das, Adam DI Kramer, Laura A Dabbish, and Jason I Hong. Increasing security sensitivity with social proof: A large-scale experimental confirmation. In *Proc. CCS*, 2014.

[7] Serge Egelman, Marian Harbach, and Eyal Peer. Behavior ever follows intention? A validation of the Security Behavior Intentions Scale (SeBIS). In *Proc. CHI*, 2016.

[8] Serge Egelman and Eyal Peer. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proc. CHI*, 2015.

[9] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. Which privacy and security attributes most impact consumers' risk perception and willingness to purchase IoT devices? In *Proc. S&P*, 2021.

[10] Nina Gerber, Benjamin Reinheimer, and Melanie Volkamer. Investigating people's privacy risk perception. In *Proc. PETS*, 2019.

[11] Matthew Hughes. Study shows we're spending an insane amount of time online. https://tinyurl.com/TimeOnline2019, 2019. Online; last accessed on 2022-05-25.

[12] Ponnurangam Kumaraguru and Lorrie Faith Cranor. *Privacy indexes: A survey of Westin's studies*. Technical Report CMU-ISRI-5-138, Carnegie Mellon University, 2005.

[13] Naresh K Malhotra, Sung S Kim, and James Agarwal. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information systems research*, 15(4):336–355, 2004.

[14] Mary L McHugh. Interrater reliability: The kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[15] William Melicher, Mahmood Sharif, Joshua Tan, Lujo Bauer, Mihai Christodorescu, and Pedro Giovanni Leon. (Do not) track me sometimes: Users' contextual preferences for web tracking. In *Proc. PETS*, 2016.

[16] Vahid Nassiri, Anikó Lovik, Geert Molenberghs, and Geert Verbeke. On using multiple imputation for exploratory factor analysis of incomplete data. *Behavior research methods*, pages 1–15, 2018.

[17] Richard G Netemeyer, William O Bearden, and Subhash Sharma. *Scaling procedures: Issues and applications*. sage publications, 2003.

[18] Elissa M. Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L. Mazurek. A comprehensive quality evaluation of security and privacy advice on the web. In *USENIX Security*, 2020.

[19] William Reynolds. Development of reliable and valid short forms of the Marlow–Crowne social desirability scale. *Journal of Clinical Psychology*, 38:119–125, 01 1982.

[20] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-confidence trumps knowledge: A cross-cultural study of security behavior. In *Proc. CHI*, 2017.

[21] Mahmood Sharif, Jumpei Urakawa, Nicolas Christin, Ayumu Kubota, and Akira Yamada. Predicting impending exposure to malicious content from user behavior. In *Proc. CCS*, 2018.

[22] Joshua Tan, Mahmood Sharif, Sruti Bhagavatula, Matthias Beckerle, Michelle L. Mazurek, and Lujo Bauer. Comparing hypothetical and realistic privacy valuations. In *Proc. WPES*, 2018.

[23] Jason Watson, Heather Richter Lipford, and Andrew Besmer. Mapping user preference to privacy default settings. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 22(6):1–20, 2015.

[24] Allison Woodruff, Vasyl Pihur, Sunny Consolvo, Laura Brandimarte, and Alessandro Acquisti. Would a privacy fundamentalist sell their DNA for $1000... if nothing bad happened as a result? The Westin categories, behavioral intentions, and consequences. In *Proc. SOUPS*, 2014.