# A High Coverage Cybersecurity Scale Predictive of User Behavior

Yukiko Sawaya
*KDDI Research, Inc.*

Sarah Lu
*MIT*

Takamasa Isohara
*KDDI Research, Inc.*

Mahmood Sharif
*Tel Aviv University*

## Abstract

Psychometric security 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 more comprehensive scales measuring a wider range of security-related factors. To test this hypothesis, we ran a series of four online studies with a total of 1,471 participants. First, we developed the extended security behavior scale (ESBS), a high-coverage scale containing substantially more items than prior ones, and collected responses to characterize its underlying structure. Then, we conducted a follow-up study to confirm ESBS's structural validity and reliability. Finally, over the course of two studies, we elicited user responses to our scale and prior ones while measuring three security behaviors reflected by Internet browser data. Then, we constructed predictive machine-learning models and found that ESBS can predict these behaviors with statistically significantly higher accuracy than prior scales (6.17%–8.53% ROC AUC), thus supporting 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 [21]. 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.

Researchers have attempted to develop a rigorous understanding of users' privacy attitudes, behaviors, concerns, and preferences to help design systems that serve users best. Among different approaches, psychometric scales such as the Privacy Concerns Scale (PCS) [6], Internet Users' Information Privacy Concerns (IUIPC) [27], the Westin Index [25], and the Security Behavior Intentions Scale (SeBIS) [11] have been proposed as affordable and scalable means to learn about users' security and privacy attitudes, concerns, and behaviors. 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 policies on social media [50]); bootstrapping personalized defenses to usable states (e.g., ones to automatically enable or block tracking per user preferences [30]); raising users' privacy awareness (e.g., when they underestimate certain risks [15]); or measuring changes in users' behaviors and concerns over time (e.g., due to interventions such as user education or the implementation of new security and privacy features [9, 12]).

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 [51]. Similarly, Tan et al. 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) [48].

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 [10]. For example, they showed that study participants who scored highly on proactive awareness were less likely to be deceived by phishing than others. Still, the correlation was far from being perfect. Furthermore, prior work found that while SeBIS responses could predict users' exposure to malicious websites, they were significantly less accurate than behavioral features at doing so ($\sim$20% lower area under the receiver operating characteristic curve) [43]. Thus, while scales such as SeBIS constitute easy-to-use instruments to learn about users, there remains significant room for improving their accuracy at predicting actual behavior.

We hypothesize that the omission of critical factors that may directly impact users' security behavior may harm scales' ability to predict actual behavior. Said differently, we expect that *scales that cover a more comprehensive set of security-related factors can predict users' behavior more accurately*. To test this hypothesis, we developed the *extended security behavior scale (ESBS)*, a scale covering a wider range of items and factors than SeBIS. To develop ESBS, we started with a comprehensive list of 374 security advice items [38], of which we selected 45 to construct the initial questions according to a well-defined, objective criteria. Following standard practices [7], we recruited 299 participants to refine the scale and discover its underlying structure. Subsequently, we remained with a scale consisting of 20 questions and five sub-scales (i.e., factors), measuring behavior related to data securement, proactive awareness, anti-virus usage, updating behavior, and password creation. In a follow-up study, we recruited another batch of 500 participants and validated the structure and reliability of ESBS. Besides being more comprehensive than SeBIS (e.g., containing 20 vs. 16 items and five vs. four factors), ESBS also allows respondents to select N/A as answers with the aim of eliciting more detailed information from them and facilitating more accurate predictions.

Following scale construction, we tested how accurately ESBS can predict users' security behavior and contrasted it with SeBIS. In particular, over the course of two studies, we recruited 672 participants, collecting their browsing history and user-agent information, and measured three behaviors: *(1)* whether they clear their browsing history;*(2)* whether they compartmentalize their browsing activities across browsers or browser profiles; and *(3)* whether they keep their operating system up-to-date. Following standard psychometric techniques—namely, factor analysis [7] and item response theory [13]—we scored ESBS by linearly combining sub-scale responses while giving equal weights to items (as in SeBIS), or by combining sub-scales responses non-linearly, scoring items per the mean and variance in participant responses. Then, we trained and evaluated predictive models relying on scale scores, the number of N/As, and demographics as independent variables. For all three security behaviors, we found that ESBS scored linearly leads to more accurate predictions than SeBIS (6.17%–8.53% higher mean receiver operating characteristic area under curve) and that accounting for N/As can help improve prediction accuracy.

In a nutshell, this work makes the following contributions:

- We propose ESBS, a high-coverage psychometric security scale containing more items and factors than SeBIS, and validate its structure and reliability.
- We found that ESBS can predict three different security behaviors more accurately than SeBIS.
- We found that including N/As can aid in improving prediction accuracy, demonstrating the utility of including them in psychometric scales as potential responses.
- We discovered that simple, linear scale scoring surpasses elaborate non-linear scoring in prediction accuracy.

The paper proceeds as follows. Next, we cover related work and background (§2) and introduce an overview of our methodology (§3). Then, we present ESBS's construction and validation processes and results (§4) before we demonstrate its improved behavior-prediction capacity (§5). We close the paper with a discussion (§6) and a conclusion (§7).

## 2 Background and Related Work

This section will present standard scale-development procedures, existing security and privacy scales, and work on security and privacy predictive analytics.

### 2.1 Psychometric Scale Development

Of many scale-development procedures proposed in prior work (e.g., [4, 7, 18, 31, 33]), the processes defined by Carpenter [7] and Netemeyer et al. [33] are of the most widely employed. At a high level, the processes start by composing a list of concrete statements, also known as scale items, to be used for capturing and measuring the different aspects of latent concepts of interest (e.g., privacy attitudes or security intentions). Then, *exploratory factor analysis (EFA)* is typically used to assess correlations between statements, and identify the structure of underlying latent constructs (i.e., sub-scales) behind the main concept. After this step, we usually remain with a subset of initial statements, each assigned to a different sub-scale. Finally, *confirmatory factor analysis (CFA)* is conducted to confirm the structure identified by EFA on a distinct sample of responses. In line with prior work (e.g., [11]), we follow a similar procedure to construct our scale.

Traditional approaches relying on factor analysis leverage linear scaling to measure humans on different sub-scales. In other words, a sub-scale identified via factor analysis is usually scored via the sum or average of responses to its corresponding statements. However, such approach does not account for the varied difficulties of items (i.e., mean response) and the variance between respondents. To this end, certain psychometric approaches, such as the notable *item response theory (IRT)* [13], aim to account of inter-statement differences, scoring respondents differently on each statement, depending on its difficulty and the variance within the population. However, IRT models are rarely used to score scales, potentially due to their complexity and limited support by open-source tools [22]. In this work, we evaluate whether (non-linear) IRT-based scaling can enable more accurate behavior predictions.

### 2.2 Security and Privacy Scales

Researchers have designed several privacy scales in past decades (e.g., [6,17,25,27]). Considered as one of the longest-

standing scales, the Westin Index helps categorize respondents as privacy fundamentalists, pragmatists, or unconcerned depending on their level of privacy concerns reflected by answers to three statements [25]. The Internet Users' Information Privacy Concerns (IUIPC) scale contains three sub-scales—control, awareness, and collection—for gauging users' privacy concerns [27]. Lastly, the Privacy Concerns Scale (PCS) is a one-dimensional scale commonly used for measuring privacy attitudes [6]. Differently from these efforts, we aim to develop a security scale geared toward assessing and predicting security-related behavior.

By contrast to privacy scales, the development of psychometric security scales mostly dates back to the last decade (e.g., [10, 14, 34, 41, 46, 49]). Notably, to our knowledge, the Security Behavior Intentions Scale (SeBIS) was the first composite scales aiming to assess user intentions to adopt a range recommended security behaviors [10]. SeBIS contains 16 items and four sub-scales gauging intentions to secure devices (e.g., via locking), generate strong passwords, update software, and proactively avoiding risks through awareness (e.g., via verifying links before clicking on them). Faklaris et al. proposed SA-6, a one-dimensional six-item scale, for assessing users' security attitudes [14]. While it seeks to measure attitudes rather than behavior intentions, Faklaris et al. found that SA-6 strongly correlates with SeBIS. In this work, we compare our scale with SeBIS, as it aims to gauge behavior more directly and is more comprehensive than alternatives.

## 2.3 Predictive Analytics

Researchers working on security and privacy predictive analytics have demonstrated the feasibility of forecasting security behavior and incidents through observations and measurements made earlier in time (e.g., [26, 40, 43, 45]). Many predictive analytics systems rely on system logs and other costly, time-consuming, and potentially invasive system measurements to enable predictions. For instance, Soska and Christin collected information about content-management system versions and installed plugins to forecast whether web servers will be compromised within a year from data collection [45]. As another example, Sharif et al. gathered HTTP logs and related Internet browsing data (e.g., website categories and amount of bytes downloaded or uploaded) to predict user exposure to malicious content [43].

Arguably, scales constitute one of the simplest and least invasive means to acquire information for predictive analytics. Indeed, prior work has shown responses to scales are somewhat predictive of actual user behavior. For instance, Egelman et al. have shown that study participants who scored higher on SeBIS's proactive awareness, updating, device securement, and password generation sub-scales also tended to detect phishing attempts more successfully, applied software updates in a more timely manner, used smartphone-locking mechanisms more often, and generated harder-to-guess passwords, respectively. Still, users' scale scores often do not correlate with their real-world behavior or leave significant room for improvement. For instance, facing the so-called privacy-paradox phenomenon, Woodruff et al. found that participants reporting high privacy concerns on the Westin Index (i.e., privacy fundamentalists) were roughly as likely as others to share sensitive data about themselves (e.g., medical records) for a certain fee [51]. Mayer et al. found that SA-6 scores were not predictive of password-meter adoption [28], while Smullen et al. discovered they were unable to forecast whether users will opt-out from online tracking [44]. Tan et al. reported 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) [48]. Similarly, Sharif et al. found that although SeBIS responses could predict users' exposure to malicious websites with certain accuracy, they were significantly less accurate than behavioral features at doing so ($\sim$20% lower area under the receiver operating characteristic (ROC) curve) [43]. This work explores the possibility of enhancing security scales' accuracy at predicting user behavior by making them more comprehensive.

# 3 Methodology Overview

Here we provide an overview of our methodology before we lay out the details and the results in the following sections. Our methodology consists of three primary stages, involving scale development, scale validation, and user-behavior prediction. In each of the first two stages we ran an online user study, while in the last stage we administered two studies, resulting in four studies overall. To avoid learning effects and ensure internal validity, we recruited independent samples across the four studies. Fig. 1 summarizes our methodology.

***Scale Development*** We initiated our work by composing the initial items to be included in our scale. We composed the items' statements by selecting appropriate security advice from Redmiles et al.'s comprehensive set of 374 advice items [38]. In particular, we selected advice that was deemed accurate, useful, of high priority, and widely relevant, as elaborated in §4.1. For each resulting item, scale respondents were given the option to respond on a 5-point Likert scale denoting the frequency in which they follow the advice, or alternately select N/A, in case they felt the statement did not apply to them or they did not understand the statement.

After composing the initial version of our scale, we conducted an online user study to refine it and identify its latent structure. We designed and ran an online user study to collect responses to the initial scale items. After collecting the responses, we removed unhelpful items (e.g., ones exhibiting ceiling or floor effects) and ran EFA to identify the sub-scales forming our primary scale, ending with $\text{ESBS}_{FA}$, a security-behavior scale containing five sub-scales and 20

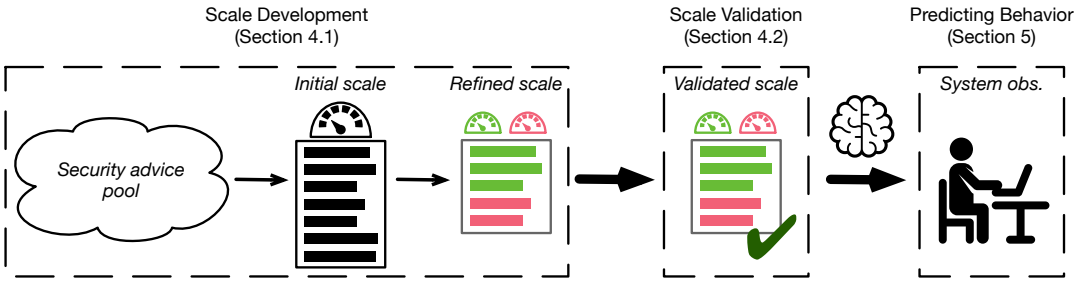


Figure 1: A high-level overview of our methodology. We started by developing the scale, first compiling an initial scale by selecting and rephrasing appropriate security advice, collecting responses via an online study, and refining the scale by running EFA to identify the sub-scales that emerge and select the items to maintain. Subsequently, we collected additional responses and validate the scale's structure (via CFA) and validity. Lastly, we ran studies, this time also collecting system-level behavioral observations, and built machine-learning models to measure the scale's accuracy when predicting behavior.

items. Additionally, we analyzed the responses with an adequate IRT model, and ended with $\text{ESBS}_{IRT}$, a scale with five equivalent sub-scales based on the same 20 items as $\text{ESBS}_{FA}$ but is scored non-linearly.

***Scale Validation*** Next, we ran another online study to assess the scale's reliability and validity. High reliability indicates that the scale and its corresponding sub-scales' items measure the same constructs, whereas high validity denotes that the same underlying construct holds on independently collected samples. After collecting another sample of responses, we ran standard reliability and validity tests and ensured that scale and the underlying structure we identified in the first stage were reliable and valid.

***Predicting Behavior*** Finally, we conducted two online studies to measure whether ESBS can predict security behavior and put out hypothesis—i.e., scales with higher coverage are able to predict behavior more accurately—to the test. In these studies, besides collecting responses to ESBS's and SeBIS's items and demographic questions, we also collected participants' browsing history or their user agent strings, to gauge whether they follow recommended security behaviors (namely, isolating browsing sessions across profiles or browsers, clearing browsing history, and frequently applying operating-system updates). Subsequently, we trained machine-learning models to predict behavior based on scale scores and demographics, and compared the prediction performance across scales and scoring techniques (i.e., via FA or IRT).

All of our user studies have been reviewed and approved by our Institutional Review Board (IRB).

# 4 Higher Coverage Security Scale

This section presents the process we followed to develop ESBS, our scale with higher coverage, and the process for refining and validating it. We also contrast ESBS with SeBIS to highlight their similarities and differences.

## 4.1 Scale Development

### 4.1.1 Methodology

***Composing Initial Items*** Similarly to SeBIS [11], 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 on Redmiles et al.'s work [38] 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 antiviruses. 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 per-

ceived risk reduction was ≥40%—the median perceived reduction among all advice.

3. **High priority**: We excluded low-priority advice items, only keeping items in the 50th percentile.

4. **Security relevance and wide applicability**: Two researchers manually and independently examined the remaining advice items and the documents originally presenting them to identify ones that are *1)* related to personal security (e.g., excluding items concerning giving advice to others); and *2)* applicable to a wide range of users (e.g., not only ones raising kids or employed by tech enterprises). Overall, they found 30 items that do not satisfy the criteria, narrowing down the advice-item list to 83. The coders had substantial inter-coder agreement (Cohen $\kappa$=0.69) [29], and resolved disagreements manually by discussing and agreeing 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.

Applying criteria 1–3 resulted in a marked decrease in the number of advice items, with 103 items surviving the selection process. 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)*. Additionally, we allow respondents to select N/A to indicate they do not understand the statement or feel it does not apply to them. When scoring the respondents, we treat N/As as never, since the respondent do not follow the describe behavior in practice. Moreover, in our experiments, we evaluate whether the additional information gained from N/As (e.g., that the participant lacks basic understanding regarding a certain behavior) can be leveraged to enhance prediction accuracy. Tab. 6 in App. B presents the initial 45 items.

***Refinement and Sub-scale Identification*** Following the recommended steps for scale development of Carpenter [7] and Netemeyer et al. [33], and SeBIS's development procedure, we designed an online user study to collect responses on our initial scale, refine it, and explore its latent construct via EFA. We asked participants in our study to fill out a survey composed of three primary parts. The first part presented the initial scale's questions, assessing the participants' propensity to follow the 45 recommended advice items initially selected. We were concerned that participants' responses were influenced by their desire to appear more socially acceptable [8]. 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 [8, 39]. Finally, we asked participants basic demographic questions. The detailed study protocol is presented in App. A.1.

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 second 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, following Egelman and Peer [11], we began the study with attention questions, notified participants that failed them once, and precluded participants that failed them twice from completing the study.

***Data Analysis*** We followed standard processes to analyze the collected data [7]. First, we examined how often questions received N/A responses. After noticing the no question stood out as widely non-applicable, we mapped N/A responses to "Never" (i.e., 1), following the intuition that if a respondent reports security advice as non-applicable, then they never follow it. As an alternative approach, we also ran an identical analysis while employing the so-called imputations technique—treating N/As as missing values, and replacing them with most likely estimates based on other participant responses [32]. This approach led to consistent findings.

Secondly, we removed items exhibiting ceiling ($\mu$>4.0) or floor ($\mu$<2.0) effects, or low variance ($\sigma$<1.0), as these have low utility in a scale due to not differentiating between respondents. We also removed items that did not exhibit high Spearman correlation (≥0.3) with other items, as they fail to measure the same construct as other items [7].

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 then tested whether all statistics lay within the recommended ranges [7].

Then, we performed EFA, using maximum-likelihood estimation and Oblimin rotation [7]. In line with prior work [19, 36], we followed an iterative EFA procedure. Specifically, after selecting the number of factors via the standard eigenvalue criterion (i.e., factors whose eigenvalue are above one) [7], we examined per-item factor loadings and removed items whose largest loadings were low (<0.32). The rationale behind removing those items is that they are unable to reliably measure a single latent construct [7]. We repeated this process until no items were removed, remaining with the refined scale and corresponding sub-scales.

Lastly, we analyzed the participants' responses using a multidimensional IRT model called the multidimensional graded response model (MGRM), a model applicable for analyzing polytomously scored items (e.g., on a Likert scale, akin to ours) and discovering latent dimensions (i.e., sub-scales) [24].

MGRM assumes that each person has a certain ability that cannot be observed directly, but can be estimated from their observed responses to a psychometric scale. The MGRM model assigns each item $k$-1 thresholds where $k$ is the number of response levels per item (e.g., $k$=5 for a 5-point Likert scale). The $i^{\text{th}}$ threshold denotes the ability score the respondent need to surpass to increase their likelihood of selecting the $i+1^{\text{th}}$ response level or higher for the given item. Moreover, MGRM assigns each item a discrimination parameter per dimension. The higher is the discrimination score, the more the item varies between individuals as their ability changes along the dimension. We estimated the MGRM model's parameters with an increasing number of dimensions, selecting the model with the lowest Bayesian information criterion (BIC) as long as increasing the number of dimensions (i.e., model complexity) markedly decreased the BIC (by $>$5), as is standard [42]. Once the model's threshold and discrimination parameters were found, we could use them to estimate the participants' latent ability scores per sub-scale from their responses through maximum *a posteriori* likelihood estimation. We used the GIRTH Python package[1] for parameter estimation.

#### 4.1.2 Participants

We recruited participants via Prolific,[2] 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 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 $>$5:1 response-to-item ratio, as recommended [7]. The average participant's age was 45.6 ($\pm$16.0) years and 47.8% of the participants reported themselves as males. It took participants an average of 9.3 minutes to complete the survey, and they were compensated 1.6 GBP ($\approx$2.0 USD) for participating.

#### 4.1.3 Results

The number of N/A responses received for each of the 45 initial items was relatively low, leading us to initially keep all items. However, we later excluded 19 items due to ceiling effects (10), low variance (11), or low total-item correlation (2), leaving us with 26 items. Tab. 6 presents detailed statistics for all items. For the remaining items, we found that Bartlett's test of sphericity ($\chi^2$=2730.0, $p$-value<0.001), Kaiser-Meyer-Olkin test of sampling adequacy (0.91), and the inter-item correlation values (0.28 on average), all laid in the recommended ranges [7, 35]. We highlight that none of the items, including those removed, were correlated with the social desirability measures, indicating that participants' responses were not influenced by a social desirability bias.

[1] https://eribean.github.io/girth/
[2] https://prolific.co

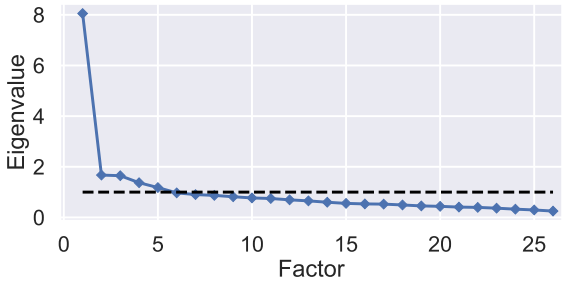


Figure 2: Eigenvalues attained from EFA. We selected five factors, as the eigenvalues of remaining factors were $<$1.

For EFA, we selected five factors according to the eigenvalue criterion (Fig. 2) [7]. We ran EFA for three iterations, removing a total of six items in the first two iterations, and none in the third. Hence, overall, we remained with 20 items in our scale. Tab. 1 presents the factors and their corresponding items, while Tab. 2 lists the factor loadings on each factor, per item. Two researchers met and named the sub-scales according to the themes of their respective items. The first factor (DS) gauges behaviors related to data securement (e.g., validating the digital certificates of HTTPS websites); the second factor (PA) assesses proactive awareness (e.g., verifying the extensions of downloaded files); the third factor (AV) measures whether individuals install and use security programs (namely, anti-viruses); the fourth factor (UP) elicits whether individuals keep their devices up-to-date; and the last factor (PW) checks whether they select strong and unique passwords. We refer to each of these identified sub-scales by $\text{ESBS}^{*}_{FA}$, where $*$ stands to the sub-scale's name (e.g., $\text{ESBS}^{DS}_{FA}$ to the data securement sub-scale identified via factor analysis and scored via summing the relevant items' responses).

The IRT analysis also resulted in five dimensions, all of which heavily align with the factors found by EFA. Tab. 2 presents the dimensions and items' discrimination-score estimates. It can be immediately seen that items typically contribute the most to the dimension corresponding to the related factor from EFA. In other words, the highest factor loadings and discrimination scores of most items are assigned to the same dimensions. Accordingly, we refer to the IRT sub-scales by $\text{ESBS}^{*}_{IRT}$, where $*$ stands for the related factor's name. A chief difference between the IRT and EFA models is that, in the case of EFA, the sub-scale scores are computed by linearly combining the scores of specific sets of items via summation, whereas, in the case of IRT, all items contribute non-linearly to all sub-scale scores (a.k.a., the ability scores), and the item contributions to each ability score are non-uniform, varying based on the thresholds and discrimination parameters.

| # | Factor 1: Data Securement (DS; 15.0% of variance explained; λ=3.00) | $\mu$ | 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 physically destroy drives I am done using and wish to erase | 2.60 | 1.68 |
| 1.6 | I encrypt my devices' disks to keep my data confidential | 2.38 | 1.48 |
| 1.7 | I safely store my private key for email encryption | 2.60 | 1.68 |
| # | **Factor 2: Proactive Awareness (PA; 11.0% of variance explained; λ=2.19)** | $\mu$ | 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 | When possible, I use two- or multi-factor authentication | 3.83 | 1.11 |
| 2.6 | I disable auto-run to prevent potentially malicious downloaded programs from running | 3.61 | 1.55 |
| # | **Factor 3: Anti-virus (AV; 7.7% of variance explained; λ=1.55)** | $\mu$ | σ |
| 3.1 | I scan attachments for viruses before downloading or opening them | 3.51 | 1.43 |
| 3.2 | I verify that my anti-virus software is up-to-date | 3.77 | 1.30 |
| 3.3 | I install anti-virus software when setting up my devices | 3.91 | 1.31 |
| # | **Factor 4: Updates (UP; 5.9% of variance explained; λ=1.17)** | $\mu$ | σ |
| 4.1 | I turn on automatic updates for devices and applications upon installation | 3.56 | 1.22 |
| 4.2 | When I am prompted about a device or software update, I immediately install it | 3.30 | 1.13 |
| # | **Factor 5: Passwords (PW; 3.8% of variance explained; λ=0.77)** | $\mu$ | σ |
| 5.1 | I select hard-to-guess passwords (with multiple character types, without dictionary words, etc.) | 3.99 | 1.11 |
| 5.2 | I select different passwords for different accounts and devices | 3.93 | 1.09 |

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

## 4.2 Scale Validation

### 4.2.1 Methodology

***Study Design*** We validated the scale by eliciting responses on the ESBS from new participants. This time, we asked the participants to respond to the ESBS's questions followed by demographics questions. Again, to ensure data quality we introduced an attention question, warned participants who failed it once, and disqualified those who erred twice. We also randomly ordered the ESBS's items to prevent ordering effects. Moreover, we advertised the study as one exploring technology perceptions to alleviate selection bias. The complete study protocol is available in App. A.2.

***Data Analysis*** We validated ESBS by testing its overall reliability, as well as the reliability of each of its sub-scales. Additionally, we tested the validity of the structure discovered in §4.1.3—i.e., checking whether the same latent structure generalizes across samples. To assess reliability, we used Cronbach's α, and checked whether it lays in the desirable range of $\geq$0.6 [3]. To check structural validity, we used a battery of goodness of fit measures and tested whether they fall within their respective recommended ranges. Specifically, we tested that the Root Mean Square Error of Approximation (RMSEA) falls below the recommended cutoff of 0.06 [20]; the Standardized Root Mean Square Residual (SRMR) lays below the suggested 0.08 [20]; the Comparative Fit Index (CFI) is above the recommended 0.9 [33]; and the Tucker-Lewis Index (TLI) falls above 0.9 [33].

### 4.2.2 Participants

We recruited participants via Prolific. We opened our study to participants from the United States who are at least 18 years old. Additionally, we collected data from a population-representative sample, resulting in a sample whose ethnicity, gender, and age distribution reflects the general United States population's. We also ensured the participants were distinct of those recruited for the first study. A total of 509 partici-

| # | EFA | | | | | IRT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | F2 | F3 | F4 | F5 | D1 | D2 | D3 | D4 | D5 |
| 1.1 | **0.755** | -0.078 | 0.100 | -0.032 | 0.049 | **1.445** | 0.742 | 1.014 | 0.021 | -0.446 |
| 1.2 | **0.585** | 0.201 | 0.098 | 0.009 | -0.167 | 1.135 | **1.301** | 0.824 | -0.039 | -0.433 |
| 1.3 | **0.625** | 0.144 | -0.035 | 0.102 | -0.167 | **1.566** | 1.501 | 1.258 | -0.261 | -0.503 |
| 1.4 | **0.572** | 0.256 | 0.081 | -0.001 | -0.205 | **1.333** | 1.186 | 1.016 | -0.023 | -0.377 |
| 1.5 | **0.400** | 0.067 | 0.048 | -0.146 | 0.163 | **1.281** | 1.110 | 0.840 | 0.044 | -0.490 |
| 1.6 | **0.721** | -0.049 | -0.011 | 0.049 | 0.214 | **1.714** | 0.961 | 1.084 | -0.093 | -0.525 |
| 1.7 | **0.762** | -0.045 | -0.045 | 0.018 | 0.050 | 0.642 | **0.846** | 0.302 | 0.287 | -0.055 |
| 2.1 | 0.097 | **0.469** | -0.022 | 0.036 | -0.068 | 0.229 | **0.850** | 0.353 | -0.078 | 0.223 |
| 2.2 | 0.056 | **0.704** | 0.122 | -0.073 | -0.055 | 0.168 | **1.673** | 0.120 | 0.005 | 0.206 |
| 2.3 | -0.02 | **0.665** | 0.015 | 0.007 | 0.130 | 0.166 | **2.040** | -0.065 | -0.047 | 0.327 |
| 2.4 | -0.068 | **0.549** | -0.061 | 0.137 | 0.053 | -0.048 | **1.131** | 0.195 | 0.259 | -0.061 |
| 2.5 | 0.187 | **0.444** | -0.032 | 0.078 | 0.185 | **0.728** | 0.637 | 0.001 | -0.001 | -0.001 |
| 2.6 | 0.160 | **0.430** | 0.087 | -0.036 | 0.053 | -0.001 | **1.710** | 0.001 | 0.001 | 0.001 |
| 3.1 | 0.128 | 0.262 | **0.336** | -0.006 | 0.150 | 0.185 | **1.474** | 1.028 | 0.413 | 0.239 |
| 3.2 | -0.026 | 0.057 | **0.875** | 0.033 | -0.041 | 0.414 | 1.896 | **2.377** | 2.140 | 1.192 |
| 3.3 | 0.035 | -0.100 | **0.761** | 0.02 | 0.074 | 0.071 | 1.468 | **2.048** | 1.776 | 0.950 |
| 4.1 | 0.024 | -0.047 | 0.075 | **0.808** | 0.009 | 0.992 | 0.485 | -0.449 | **2.260** | -0.296 |
| 4.2 | -0.020 | 0.068 | -0.054 | **0.672** | -0.015 | 0.602 | 0.514 | -0.366 | **1.369** | 0.105 |
| 5.1 | 0.033 | 0.331 | 0.118 | -0.008 | **0.486** | 0.901 | 0.825 | 0.337 | 0.001 | **1.261** |
| 5.2 | 0.120 | 0.115 | 0.090 | -0.017 | **0.510** | 1.929 | 1.325 | -0.001 | -0.001 | **2.597** |
| α | 0.857 | 0.766 | 0.760 | 0.704 | 0.614 | - | - | - | - | - |

Table 2: A summary of the EFA and IRT results. For EFA, we report the per-item factor loadings on the identified five factors. We also report Cronbach's Alpha (α) for each factor. The scale's overall α is 0.882. For IRT, we report the discrimination values per dimension for each of the items. The maximum factor loading and discrimination value for each item are in boldface.

pants started the study, and nine dropped out due to failing the attention question or quitting our study. Thus, overall, 500 participants completed our study. The average participant's age was 45.4 (±16.3) years and 47.6% of the participants reported themselves as males. It took participants an average of 5.1 minutes to complete the survey, and they were compensated 0.82 GBP (≈1.0 USD) for participating.

#### 4.2.3 Results

Tab. 3 presents the reliability and validity metrics we calculated. The results show that all metrics fall within the ranges recommended in the literature. Hence, we can conclude that ESBS and each of its sub-scales are reliable, and that the scale's structural validity holds.

| | Metric | Value | Recommended |
|---|---|---|---|
| Reliability | α(DS) | **0.83** | ≥0.60 |
| | α(PA) | **0.71** | |
| | α(AV) | **0.80** | |
| | α(UP) | **0.64** | |
| | α(PW) | **0.71** | |
| | α(Overall) | **0.88** | |
| Validity | CFI | **0.93** | ≥0.90 |
| | TLI | **0.91** | ≥0.90 |
| | RMSEA | **0.05** | ≤0.06 |
| | SRMR | **0.05** | ≤0.08 |

Table 3: Reliability and validity measures and their recommended ranges. All lay within recommended ranges.

### 4.3 Comparison With SeBIS

It is instructive to contrast ESBS with SeBIS. Two members of the research independently coded each of the SeBIS and ESBS items according to whether they are included in the other scale. The researchers agreed on 13 of the 16 SeBIS items and resolved differences by accepting one of the researchers' codes. Eight of the SeBIS items were not included in ESBS due to exhibiting ceiling effects of low factor loadings. Of the eight SeBIS items that were included in ESBS, two overlapped with a single ESBS item. Hence, 13 ESBS

items were not covered by SeBIS. Specifically, these included ESBS items 1.1, 1.3–1.7, 2.1, 2.3–2.6, 3.1, and 3.3 (see Tab. 1). These items cover a wide range of critical security behavior, including the encryption and protection of sensitive information users send online, the verification of parties they communicate with via email and messaging applications, the use of multi-factor authentication, and the deployment of anti-viruses [38]. By including these items, ESBS covers a wider range of behaviors than SeBIS. In what follows, we verify whether this higher coverage translates to higher prediction accuracy of actual user behavior.

## 5 Behavior Prediction

After developing ESBS and validating it, we ran two studies to test whether it can predict user behavior and compare its prediction accuracy with that of SeBIS. In one study, we collected Internet-browsing data and attempted to predict whether participants isolate browsing sessions across different profiles or browsers, and whether they frequently clean their browsing history. In the other study, we sought to predict whether users keep their operating systems up-to-date.

### 5.1 Browsing-Behavior Study

#### 5.1.1 Methodology

***Study Design*** We administered a user study to gather responses to ESBS and SeBIS, and collect browsing data. In this study, we presented participants with ESBS and SeBIS questionnaires. As in the studies above, additionally to the 5-point Likert scales, study participants could also respond with N/A to ESBS items. Subsequently, participants were asked to install a Chrome extension to share their browsing history with the researchers. We selected Chrome as it is the browser with the largest market share—62.85% of Internet users set Chrome as their default browser [47]. To ameliorate privacy concerns, we limited the data collection to the last three months of browsing history, collecting only first-party websites and fully qualified domains visited (instead of full URLs). Furthermore, we allowed study participants to remove domains they preferred not to share before uploading the data. For each domain shared by participants, we also included the timestamp of the visit. After completing the upload, participants were presented with demographic questions, asking about their age, gender, education, occupation, and income. In the absence of automatic means to collect data about participant behavior via the extension—i.e., test if they clear their browsing history or compartmentalize browsing sessions—we resorted to collecting self-reported data. However, we later verified that the reported data were consistent with the logged browsing histories (§5.1.3). Specifically, we asked participants whether they use unique profiles or browsers for participating in online study, separating their study-taking activities from everyday browsing. Such compartmentalization would indicate that the participants are security- and privacy-aware, as isolating browsing sessions is a recommended behavior to help circumvent online tracking and protect one's data when engaging risky activities [38]. Moreover, we asked participants about whether they have cleared their browsing history, fully or partially, within the past three months. Participants could answer affirmatively, negatively, or most likely. Again, clearing the browsing history is a recommended security and privacy behavior [38], as it could, for example, help remove login tokens that could be (mis)used by adversaries to impersonate the user. App. A.3 contains the study's protocol.

As before, in this study too we used attention questions to ensure data quality. We also presented the SeBIS and ESBS questionnaires and their respective items in random order to mitigate ordering and learning effects. To comply with the study platform's terms of service, we let participants know in advance that they will be asked to install an extension that will share browsing data with researchers. Yet, to alleviate self-selection bias, we avoided listing the study as a security study, advertising it as one aiming to measure technology perception and Internet usage. Furthermore, we limited participation from Chrome browsers, as our extension only supports uploading data from Chrome.

***Data Analysis*** We initiated the analysis by validating the participants' responses, checking that participants who self-reported to compartmentalize browsing activities indeed mostly visited study-taking platforms, while those who self-reported to have cleared their browsing history uploaded records that span shorter time periods than others.

Next, we tested and compared the behavior-prediction accuracy using SeBIS and ESBS. To this end, we built machine-learning models aiming to predict whether participants behave securely based on scale responses and demographics. More concretely, the dependent variable of the models was an indicator variable denoting whether a participant behaves securely (i.e., compartmentalize sessions or clear browsing history) or not. The independent variables primarily included the scores on each sub-scale—four scores in the case of SeBIS and five in the case of ESBS. To compare linear and non-linear scoring, we trained and evaluated models with EFA- and IRT-derived scores for both SeBIS and ESBS. Additionally, to evaluate whether N/As are conducive for predictions, we tested ESBS models with and without an additional independent variable representing the number of N/A responses entered by the participant. I.e., this resulted in six conditions compared per security behavior: linearly scored SeBIS ($\text{SeBIS}_{FA}$), non-linearly scored SeBIS ($\text{SeBIS}_{IRT}$), linearly scored ESBS with or without N/A count ($\text{ESBS}_{FA}$ and $\text{N/A+ESBS}_{FA}$), non-linearly scored ESBS with or without N/A count ($\text{ESBS}_{IRT}$ and $\text{N/A+ESBS}_{IRT}$). To account for demographics, we also included independent variables for age (in years), and indicators for higher education (Bachelor's or above), high income (75K USD or higher yearly income), gender (female or not), and

technical background. Of different machine-learning models we tested, random forests attained the highest accuracy. Thus, we ran 20 evaluation rounds (90-10 training-test splits), in which we performed a grid-search with three-fold cross-validation to choose the best model hyperparameters per condition, and evaluated performance on the test samples. As the accuracy metric, we used the receiver operating characteristic area under curve (ROC AUC), whose value ranges from 0 to 1, and tends to 1 as the model is more accurate. After finishing the evaluation, we performed a paired t-test to compare the mean AUC with the SeBIS$_{FA}$ condition.

We were also interested in assessing how each independent variable affects prediction accuracy. Therefore, we used the well-established permutation-importance model interpretation technique [5]. Given a trained model, this technique permutes the values of an independent variable across samples and measures the resulting decrease in accuracy. This process is repeated several times, permuting values of one independent variable at a time, and the average decrease in accuracy is eventually emitted as a measure of feature importance—the higher decrease in accuracy due to permutation, the more important is the independent variable for classification, as perturbing it markedly harms classification accuracy. We report the average permutation-importance scores across all evaluation rounds per feature.

#### 5.1.2 Participants

We recruited 228 participants for this study through Prolific, while limiting participation to individuals who are at least 18 years old, are located in the United States, and had not participated in the two previous studies. We recruited a gender-balanced sample, resulting in 111 respondents who identified as females. The average participant age was 40.04 ($\pm$13.12) years. It took participants roughly 12 minutes on average to complete the study. Considering the sensitivity of the browsing data we collected, we increased the compensation compared to the previous studies to 2.48 GBP ($\approx$3.0 USD), after drawing on Tan et al.'s estimate of the cost at which individuals are willing to sell their browsing history [48].

#### 5.1.3 Results

On average, each participants reported 1,279 visits to 18 distinct first-party domains. Most participants submitted their entire available browsing history for the past three months, with only 24 of the 228 participants removing and average of 9.17% of the items from their history. Browsing histories spanned at least two days for $\geq$89.47% of the participants, suggesting that the study did not trigger most participants to (fully) clear their history (arguably, participants who cleared their history before participating may be intuitively deemed security- and privacy-aware). 21.05% of the participants reported using a unique profile or browser to participate in online studies and

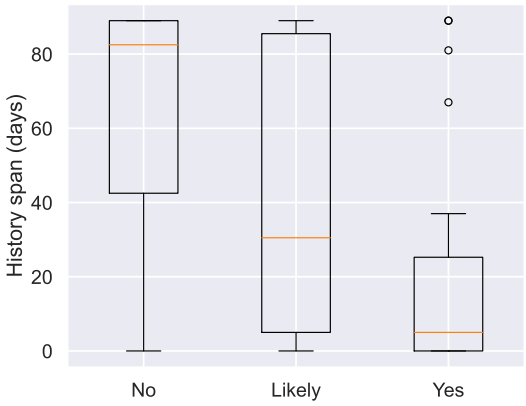


Figure 3: Boxplots depicting the time span covered by the browsing history of participants reporting to have cleared, likely cleared, or not cleared their browsing history.

crowdsourcing tasks. While we cannot fully verify compartmentalization behavior using our data alone, upon examining the domains visited most often by the participants, we found that indeed visits to crowdsourcing domains constituted an overwhelming majority of the total visits in their browsing history—a median of 71.59% of the total visits were to crowdsourcing websites. By contrast, for users reporting not using unique profiles or browsers to partake in studies, a median of only 42.54% of their total visits were to crowdsourcing domains. These self-reports were reliable, thus we used them to label participants according to whether they compartmentalize their browsing activity, separating their study-taking activities from everyday browsing. The boxplots in Fig. 3 depicts the time span covered by browsing histories of participants reporting clearing, likely clearing, or not clearing their browsing history in the three months prior to uploading their data. It can be seen that participants reporting clearing their browsing history shared histories spanning markedly shorter time spans than others, with all spanning 39 days or fewer when excluding the three outliers. Still, some participants reporting clearing or likely clearing their history shared data spanning roughly three months. Hence, when labeling our data, we avoided using potentially unreliable self-reports, and instead treated participants who shared data spanning $<$39 days as ones who clean their history. Three participants reported clearing their browsing history but shared browsing data spanning $>$39 days. We categorized those participants as ones who have not cleared their history. Labeling them as ones who have cleared their browsing history did not impact our takeaways.

Tab. 4 presents the ROC AUC mean and standard deviation per condition. The results highlight that ESBS$_{FA}$ led to higher prediction accuracy (1.70%–6.48% higher mean AUC) than SeBIS, when scores via linearly (FA) or non-linearly (IRT). These differences were statistically significant. Furthermore,

| **Dependent variable: Compartmentalization** | | |
|---|---|---|
| **Condition** | **AUC** | **(±std)** |
| SeBIS$_{FA}$ | 57.04% | (±15.35%) |
| SeBIS$_{IRT}$ | 63.87% | (±11.99%) |
| ESBS$_{FA}$ | 65.57% | (±12.33%) |
| ESBS$_{IRT}$ | 58.03% | (±12.98%) |
| N/A+ESBS$_{FA}$ | 69.60% | (±10.60%) |
| N/A+ESBS$_{IRT}$ | 61.08% | (±12.54%) |
| **Dependent variable: Clearing history** | | |
| **Condition** | **AUC** | **(±std)** |
| SeBIS$_{FA}$ | 62.57% | (±10.70%) |
| SeBIS$_{IRT}$ | 59.42% | (±10.88%) |
| ESBS$_{FA}$ | 69.05% | (±12.47%) |
| ESBS$_{IRT}$ | 63.17% | (±13.94%) |
| N/A+ESBS$_{FA}$ | 68.76% | (±11.22%) |
| N/A+ESBS$_{IRT}$ | 65.47% | (±11.91%) |

Table 4: ROC AUCs of predicting dependent variables derived from participants' browsing histories (i.e., compartmentalization and clearing history) using scale responses, demographics, and (sometimes) number of N/As. For each condition, we report the mean and standard deviation (std) of the AUC. Underlined values denote mean AUCs statistically significantly larger than those achieved by SeBIS$_{FA}$ ($p$-value$<$0.05).

incorporating the number of N/As as an independent variable either preserved prediction accuracy (for clearing history) or led to a statistically significant increase in prediction accuracy (4.03% higher AUC for compartmentalization in condition N/A+ESBS$_{FA}$ compared to ESBS$_{FA}$). Finally, it can be seen that non-linear scoring with IRT generally led to lower prediction accuracy than linear scoring with FA.

We also tested whether incorporating SeBIS$_{FA}$ scores alongside N/A+ESBS$_{FA}$ could increase prediction accuracy. The models we built using both scales as independent variables attained 66.40%±12.11% and 68.70%±11.73% mean AUC for compartmentalization and clearing history, respectively, failing to improve the accuracy achieved via N/A+ESBS$_{FA}$ alone. This finding further supports our analysis in §4.3, showing that informative SeBIS items are covered by ESBS. Hence, combining scores on both scales is unlikely to introduce additional information boosting prediction accuracy.

Figs. 4–5 present the feature importance scores assigned to the independent variables in the N/A+ESBS$_{FA}$ condition. Age was the most predictive independent variable both when predicting whether the participants compartmentalize browsing activities or clear their history, leading to a drop of roughly 5.91–7.54% in mean AUC when permuted. Closely behind are the ESBS sub-scales. When predicting compartmentalization, the data-securement (ESBS$^{DS}_{FA}$) and proactive-awareness (ESBS$^{PA}_{FA}$) sub-scales received high scores—2.55% and 2.88% reduction in AUC when permuted, respectively. This result supports the intuition that users reflecting high

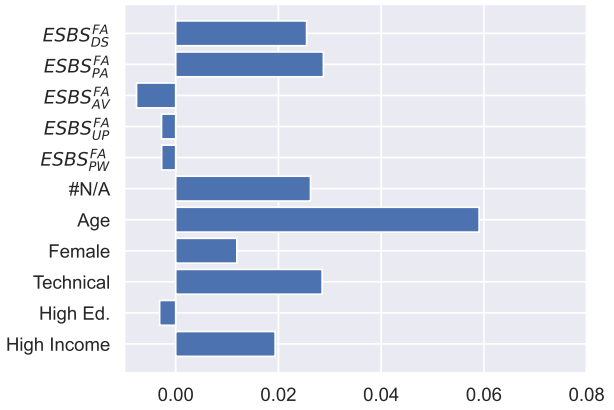


Figure 4: Mean importance of the independent variables when predicting browsing-activity compartmentalization, as calculated by the permutation test.

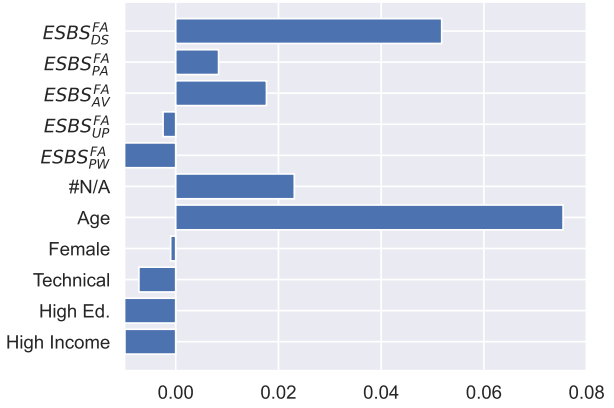


Figure 5: Mean importance of the independent variables when predicting whether users clear their browsing history, as calculated by the permutation test.

data securement and awareness intentions are more likely to compartmentalize their browsing than others. When predicting whether participants clear history, the data-securement sub-scale (ESBS$^{DS}_{FA}$) received the highest importance score amongst all sub-scales, again supporting intuition.

## 5.2 Updating-Behavior Study

### 5.2.1 Methodology

***Study Design*** In the last study, following prior work [10], and given the recent release of MacOS Ventura [23], we elected to investigate whether ESBS responses can enable accurate prediction of MacOS updates. More specifically, we sought to test whether ESBS can predict whether participants' MacOS versions are eligible for security updates. Intuitively, as participants whose MacOS version is more than three versions old do not receive security updates, they are more vulnerable to at-

tacks than others. Hence, a desirable user behavior is that they update their MacOS version to receive security updates. We limited participation in the study to MacOS users. We asked participants to answer the ESBS and SeBIS questionnaires, followed by demographic questions. Lastly, we asked participants to self-report their MacOS version from a drop-down list, giving them instructions for how to find the exact version in case they were uncertain about it. In the background, similarly to prior work [10], we also collected the user-agent string to further confirm the MacOS version. The complete study protocol is available in App. A.4. Here too, we avoided ordering effects by presenting the SeBIS and ESBS questionnaires and their respective items in a random order. We also advertised the study as a technology-perceptions study to minimize selection bias.

***Data Analysis*** Except for the validation of the self-reported MacOS versions and the derivation of the dependent variable in the machine-learning model, our analysis was similar to the previous study. Specifically, as the user-agent string can be misleading (e.g., it can be altered by certain browser extensions [16]) and participants may misreport their MacOS version, we compared self-reports with user-agent strings, removing inconsistent responses, to ensure data validity. Our dependent variable denoted whether the user's MacOS version is still eligible for security updates (i.e., version 11 or above [23]) or not.

### 5.2.2 Participants

We recruited a set of new participants, distinct of those included in the studies previously presented, via Prolific. We opened our study to participants from the United States who are at least 18 years old, using Prolific's functionality to recruit a gender-balanced sample. A total of 558 participants started the study, and 57 of them dropped out due to failing the attention question (7), quitting the study (4), or disqualifying because they did not use MacOS (46). Overall, 501 participants completed our study. The average participant's age was 36.1 ($\pm$13.2) years and 49.3% of the participants self-identified as males. It took participants 5.1 minutes on average to complete the study, and they were compensated 0.82 GBP ($\approx$1.0 USD) for participating.

### 5.2.3 Results

Of the 501 participants, the self-reports of 57 were inconsistent with the user-agents, leaving us with 444 participants to train models and evaluate prediction accuracy. Tab. 5 presents the prediction mean ROC AUC for different conditions. Here too, $\text{ESBS}_{FA}$ had statistically significantly higher mean AUC than $\text{SeBIS}_{FA}$ (+6.17% mean AUC). Non-linear scoring of SeBIS ($\text{SeBIS}_{IRT}$), increased prediction's mean AUC by 3.41% compared to linear scoring ($\text{SeBIS}_{FA}$), however, the mean AUC did not surpass that of predictions with $\text{ESBS}_{FA}$.

| **Dependent variable: MacOS updating** | | |
|---|---|---|
| **Condition** | **AUC** | **($\pm$std)** |
| $\text{SeBIS}_{FA}$ | 51.13% | ($\pm$10.59%) |
| $\text{SeBIS}_{IRT}$ | 54.54% | ($\pm$8.75%) |
| $\text{ESBS}_{FA}$ | <u>57.30%</u> | ($\pm$9.82%) |
| $\text{ESBS}_{IRT}$ | 54.50% | ($\pm$9.86%) |
| $\text{N/A+ESBS}_{FA}$ | <u>56.09%</u> | ($\pm$12.51%) |
| $\text{N/A+ESBS}_{IRT}$ | 54.15% | ($\pm$12.39%) |

Table 5: ROC AUCs of predicting MacOS updates using responses, demographics, and (sometimes) number of N/As. For each condition, we report the mean and standard deviation (std) of the AUC. Underlined values denote mean AUCs statistically significantly larger than those achieved by $\text{SeBIS}_{FA}$ ($p$-value$<$0.05).

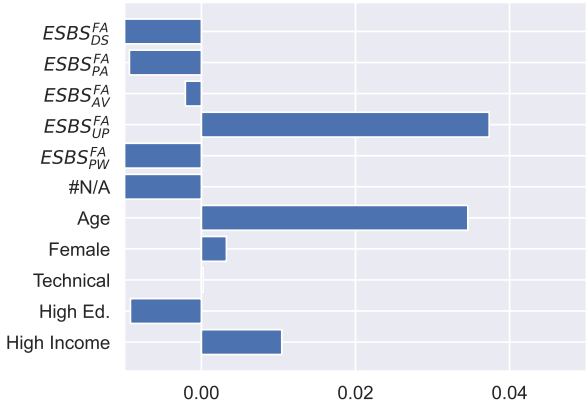


Figure 6: Mean importance of the independent variables when predicting whether participants keep their MacOS up-to-date, as calculated by the permutation test.

Incorporating the number of N/As alongside $\text{ESBS}_{FA}$ decreased the mean AUC by 1.21%, but the decrease was statistically insignificant.

Fig. 6 presents the feature importance scores assigned to the independent variables in the $\text{N/A+ESBS}_{FA}$ condition. As anticipated, the updating sub-scale ($\text{ESBS}^{UP}_{FA}$) was most predictive of whether the participants had an up-to-date MacOS version, leading to a 3.74% decrease in mean AUC when permuted. As for other security behavior, the participants' age was again a strong predictor of actual updating behavior with a 3.46% importance score.

## 6 Discussion

***Limitations*** Several limitations should be taken into consideration when interpreting our work. First, our participants, all of whom are located in the United States, were recruited through Prolific. It is well-known that Prolific users are younger and

more educated than the general public. Hence, while the participants represent a non-negligible fraction of Internet users, the extent to which our results generalize to the general Internet-user population, across cultures, ages, and education levels remains to be tested. Second, we showed that ESBS can help predict three security behaviors more accurately than SeBIS. The validation of the result across three behaviors is indeed encouraging. Nonetheless, it would be worthwhile to test whether the result generalizes across more behaviors. Particularly, behaviors related to the security of popular computing platforms not covered in this work (e.g., mobile and cloud) would be intriguing to study. Third, participants' self-reported behaviors could be inaccurate. We, however, ensured that self-reports were consistent with logged data, indicating that they were of high quality. Thus, inaccuracies, if any, are expected to be minimal. Lastly, we evaluated a finite set of machine-learning models. While we identified and used the best performing model type (i.e., random forests), it is conceivable that their might exist another model that could outperform it. Still, we believe our findings are informative, as they hold across the most commonly used models in practice.

***Means to Leverage ESBS*** Similarly to prior scales [10], ESBS can serve both researchers and practitioners. Researchers can employ ESBS as an inexpensive and non-invasive means to explore how user behavior differs across cultures [41] or evolves over time, for instance, due to changes in technological norms or user exposure to advice or education material. Practitioners, on the other hand, can exploit ESBS's predictive power to enable effective, personalized interventions. For instance, employers surveying their employees can target security advice and education material to employees anticipated to be at certain risk (e.g., due to not separating personal browsing activities from those related to work). One may also seek to leverage ESBS to tailor defenses to users. For instance, if a user is expected to avoid timely updates, aggressive interventions targeting them, such as blocking or throttling access to the employer systems, may help encourage them to update their system.

The ESBS sub-scales can be used together or separately. Although individual sub-scales may be more helpful than others for predicting specific behaviors (Figs. 4–6), we recommend using the full scale if maximizing prediction accuracy is a primary objective. For example, our experiments showed that the $\text{ESBS}_{FA}$ and N/A+$\text{ESBS}_{FA}$ conditions incur up to 5% ROC AUC loss when relying only on the most important sub-scale per Fig. 5 (i.e., data securement) instead of all sub-scales to predict whether participants clear their browsing history.

***Trade-offs Between SeBIS and ESBS*** Due to including more questions, ESBS has higher time overhead than SeBIS. A rough estimate based on our data shows it took participants ∼54 seconds on average to respond to SeBIS compared to ∼68 seconds needed for the ESBS. While the additional time imposed by ESBS could increase dropout rates, we have not observed such an increase—the incompletion rates due to quitting in the middle or failing attention checks were low (<3%) regardless of whether we asked participants to fill out only the ESBS questionnaire (§4.2) or both ESBS and SeBIS (§5). Hence, the higher predictive accuracy of ESBS may justify the minimal overhead compared to SeBIS, primarily when ESBS is employed in settings where accuracy is critical (e.g., studying user behavior over time).

***Future Work*** To maximize the utility of ESBS, it is crucial to ensure that it is widely accessible. To this end, we intend to conduct cognitive and readability studies (e.g., employing measures such as Smart Cloze [37]) to assess the scale's accessibility, compare it with prior scales, and identify means to improve it. Furthermore, we plan to run cross-cultural studies to adapt ESBS to different cultures, with a primary focus on under-studied populations that are not western, educated, industrialized, rich, and democratic (WEIRD) [41].

## 7 Conclusion

This paper offers the extended security behavior scale (ESBS), a high-coverage security scale consisting of 20 questions and five sub-scales. We developed ESBS through a rigorous process and validated its structural validity and reliability through an independent study. Furthermore, we showed that ESBS can predict a variety of security behaviors at a higher accuracy than commonly used prior scales. In particular, we measured three user behaviors, including compartmentalization of browsing activities, clearing browsing history, and updating operating systems, and showed that ESBS can predict whether users follow these behaviors with 6.17%–8.53% higher ROC AUC than SeBIS [11]. Among others, we also found that N/As can help boost prediction accuracy, and that linear scoring of scales typically leads to better predictions than elaborate non-linear scoring.

## Acknowledgments

We would like to thank the reviewers and the anonymous shepherd for their helpful feedback. This work was partially supported by a grant from the Blavatnik Interdisciplinary Cyber Research Center (ICRC); by Len Blavatnik and the Blavatnik Family foundation; by a Maof prize for outstanding young scientists; and by a gift from the Neubauer Family foundation.

## 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] Richard P Bagozzi and Youjae Yi. On the evaluation of structural equation models. *Journal of the academy of marketing science*, 16:74–94, 1988.

[4] Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health*, 6:149, 2018.

[5] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[6] 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.

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

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

[9] 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.

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

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

[12] 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.

[13] Susan E Embretson and Steven P Reise. *Item response theory*. Psychology Press, 2013.

[14] Cori Faklaris, Laura Dabbish, and Jason I. Hong. A self-report measure of end-user security attitudes (SA-6). In *Proc. SOUPS*, 2019.

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

[16] Google. User-agent switcher for Chrome. https://tinyurl.com/ChromeUASwitcher, 2021.

[17] Rakibul Hasan, Rebecca Weil, Rudolf Siegel, and Katharina Krombholz. A psychometric scale to measure individuals' value of other people's privacy (VOPP). In *Proc. CHI*, 2023.

[18] Timothy R Hinkin, J Bruce Tracey, and Cathy A Enz. Scale construction: Developing reliable and valid measurement instruments. *Journal of Hospitality & Tourism Research*, 21(1):100–120, 1997.

[19] Laura Hoffmann, Nikolai Bock, and Astrid M. Rosenthal v.d. Pütten. The peculiarities of robot embodiment (EmCorp-Scale): Development, validation and initial test of the embodiment and corporeality of artificial agents scale. In *Proc. HRI*, 2018.

[20] Li-tze Hu and Peter M Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1):1–55, 1999.

[21] Matthew Hughes. Study shows we're spending an insane amount of time online. https://tinyurl.com/TimeOnline2019, 2019. Online; last accessed on 2023-06-04.

[22] Jason C Immekus, Kate E Snyder, and Patricia A Ralston. Multidimensional item response theory for factor structure assessment in educational psychology research. *Frontiers in Education*, 4:45, 2019.

[23] Apple Inc. Apple security updates. https://support.apple.com/en-us/HT201222, 2023. Online; last accessed on 2023-05-01.

[24] Shengyu Jiang, Chun Wang, and David J Weiss. Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in psychology*, 7:109, 2016.

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

[26] Yang Liu, Armin Sarabi, Jing Zhang, Parinaz Naghizadeh, Manish Karir, Michael Bailey, and Mingyan Liu. Cloudy with a chance of breach: Forecasting cyber security incidents. In *Proc. USENIX Security*, 2015.

[27] 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.

[28] Peter Mayer, Collins W Munyendo, Michelle L Mazurek, and Adam J Aviv. Why users (don't) use password managers at a large educational institution. In *Proc. USENIX Security*, 2022.

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

[30] 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.

[31] Fabiane FR Morgado, Juliana FF Meireles, Clara M Neves, Ana Amaral, and Maria EC Ferreira. Scale development: Ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica*, 30, 2017.

[32] 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.

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

[34] Boon-Yuen Ng, Atreyi Kankanhalli, and Yunjie Calvin Xu. Studying users' computer security behavior: A health belief perspective. *Decision Support Systems*, 46(4):815–825, 2009.

[35] Ralph L. Piedmont. *Inter-item Correlations*, pages 3303–3304. 2014.

[36] Astrid Rosenthal-Von Der Pütten and Nikolai Bock. Development and validation of the self-efficacy in human-robot-interaction scale (se-hri). *J. Hum.-Robot Interact.*, 7(3), dec 2018.

[37] Elissa Redmiles, Lisa Maszkiewicz, Emily Hwang, Dhruv Kuchhal, Everest Liu, Miraida Morales, Denis Peskov, Sudha Rao, Rock Stevens, Kristina Gligorić, Sean Kross, Michelle Mazurek, and Hal Daumé III. Comparing and developing tools to measure the readability of domain-specific texts. In *Proc. EMNLP-IJCNLP*, 2019.

[38] 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 *Proc. USENIX Security*, 2020.

[39] 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.

[40] Armin Sarabi, Parinaz Naghizadeh, Yang Liu, and Mingyan Liu. Risky business: Fine-grained data breach prediction using business profiles. *Journal of Cybersecurity*, 2(1):15–28, 2016.

[41] 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.

[42] Howard J Seltman. *Experimental design and analysis*. Pittsburgh: Carnegie Mellon University, 2012.

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

[44] Daniel Smullen, Yuanyuan Yao, Yaxing Feng, Norman Sadeh, Arthur Edelstein, and Rebecca Weiss. Managing intrusive practices in the browser: A user centered perspective. In *Proc. PoPETS*, 2021.

[45] Kyle Soska and Nicolas Christin. Automatically detecting vulnerable websites before they turn malicious. In *Proc. USENIX Security*, 2014.

[46] Jeffrey M Stanton, Kathryn R Stam, Paul Mastrangelo, and Jeffrey Jolton. Analysis of end user security behaviors. *Computers & security*, 24(2):124–133, 2005.

[47] StatCounter. Browser market share worldwide. https://gs.statcounter.com/browser-market-share, 2023. Online; last accessed on 2023-06-04.

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

[49] Daniel Votipka, Desiree Abrokwa, and Michelle L Mazurek. Building and validating a scale for secure software development self-efficacy. In *Proc. CHI*, 2020.

[50] 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.

[51] 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.

# A Study Protocols

## A.1 Scale development

**A.1.1 This study requires you to voice your opinion using the scales below. It is important that you take the time to read all instructions and that you read questions carefully before you answer them. Previous research on preferences has found that some people do not take the time to read everything that is displayed in the questionnaire. The questions below serve to test whether you actually take the time to do so. Therefore, if you read this, please answer 'three' on the first question, add three to that number and use the result as the answer on the second question. Thank you for participating and taking the time to read all instructions.**

- I would prefer to live in a large city rather than a small city.
  - [Seven-levels Likert scale: Strongly disagree; Disagree; Weakly disagree; Neutral; Weekly agree; Agree; Strongly agree.]
- I would prefer to live in a city with many cultural opportunities, even if the cost of living was higher.
  - [Seven-levels Likert scale, from strongly disagree to strongly agree.]

**A.1.2 Please answer the following questions by selecting the Likert-scale option that best describes how often you follow the described behavior (i.e., Never, Rarely, Sometimes, Often, Always). Alternately, if you are unfamiliar with the described behavior or believe it is inapplicable to you, please select the N/A option.**

- [Initial 45 questions (i.e., all items); see Tab. 6 ]

**A.1.3 Listed below are a number of statements concerning personal attitudes and traits. Please read each item and decide how it pertains to you. Please respond to each item by either TRUE (T) or FALSE (F). Indicate your response by selecting the appropriate answer next to the item.**

- [Marlowe-Crowne's social desirability questions [39].]

**A.1.4 What is your age in years?**

**A.1.5 What is your gender?**

- [Options: Male; Female; Non-binary / third gender; Prefer not to say.]

**A.1.6 Which of the following best describes your highest achieved education level?**

- [Options: Some high school; High school graduate; Some college, no degree; Associates degree; Bachelor's degree; Master's degree; Doctorate (PhD, MD, or similar); Other; Prefer not to answer.]

**A.1.7 Which of the following best describes your primary occupation?**

- [Options: Administrative support (e.g., secretary, assistant); Art, writing, or journalism (e.g., author, reporter, sculptor); Business, management, or financial (e.g., manager, accountant, banker); Education or science (e.g., teacher, professor); Legal (e.g., lawyer); Medical (e.g., doctor, nurse, dentist); Computer engineer or IT professional (e.g., programmer, IT consultant); Engineer in another field (e.g., civil or bio-engineer); Service (e.g., retail clerk, server); Unemployed; Retired; College student; Graduate student; Crowdsourcing worker (e.g., Mechanical Turk worker); Prefer not to answer.]

**A.1.8 What is your household's income?**

- [Options: Less than $15,000/year; $15,000/year-$24,999/year; $25,000/year-$34,999/year; $35,000/year-$49,999/year; $50,000/year-$74,999/year; $75,000/year-$99,999/year; $100,000/year-$149,999/year; $150,000/year-$199,999/year; $200,000/year or above; Prefer not to answer.]

**A.1.9 Which device types do you use? Please select all applicable types.**

- [Options: Smartphone - Android; Smartphone - iPhone; Smartphone - Others; Laptop - MacOS; Laptop - PC/Windows; Laptop - Other; Desktop - MacOS; Desktop - PC/Windows; Desktop - Other; Tablet - Android; Tablet - iPad; Tablet - Windows; Tablet - Other.]

## A.2 Scale Validation

**A.2.1 [Attention check, same as App. A.1.1.]**

**A.2.2 Please answer the following questions by selecting the Likert-scale option that best describes how often you follow the described behavior (i.e., Never, Rarely, Sometimes, Often, Always). Alternately, if you are unfamiliar with the described behavior or believe it is inapplicable to you, please select the N/A option.**

- [ESBS questions; see Tab. 6 ]

**[Six demographic questions; same as Apps. A.1.4–A.1.9.]**

## A.3 Predicting Behavior (Browsing-Behavior)

**A.3.1 [Attention check; same as App. A.1.1.]**

**A.3.2 [ESBS question; same as App. A.2.2.]**

**A.3.3 [SeBIS questions; same as Egelman et al. [11].]**

**A.3.4 [Direct study participants to install Chrome extension and upload their browsing history.]**

**[Six demographic questions; same as Apps. A.1.4–A.1.9.]**

## A.4 Predicting Behavior (MacOS updating)

**A.4.1 [Attention check; same as App. A.1.1.]**
**A.4.2 [ESBS question; same as App. A.2.2.]**
**A.4.3 [SeBIS questions; same as Egelman et al. [11].]**
**[Six demographic questions; same as Apps. A.1.4–A.1.9.]**
**A.4.4 Which macOS version are you currently using? If unsure, please check the macOS version by selecting the Apple logo in the menu bar followed by "About This Mac."**

- [Options: Ventura (13); Monterey (12); Big Sur (11); Catalina (10.15); Mojave (10.14); High Sierra (10.13); Sierra (10.12); El Capitan (10.11); Yosemite (10.10); Mavericks (10.9); Mountain Lion (10.8); Lion (10.7); Snow Leopard (10.6); Leopard (10.5); Tiger (10.4); Panther (10.3); Jaguar (10.2); Puma (10.1); Cheetah (10.0).]

**[The user-agent string is logged in the background, without user interaction.]**

# B All Initial Items

Tab. 6 presents all 45 items initially considered for inclusion in ESBS and reports the reason why certain items were eventually excluded.

| # | Description | $\mu$ | $\sigma$ | N/As | Corr. | Removed |
|---|---|---|---|---|---|---|
| 1 | I verify that recipients are reputable before sharing sensitive information | 4.33 | 0.97 | 0.02 | - | $\mu$,$\sigma$ |
| 2 | I verify whom I communicate with online (via email or online messaging apps) is really the person I intend to | 3.92 | 1.18 | 0.03 | 2 | |
| 3 | I share my email address in return for free samples and products | 2.45 | 1.10 | 0.01 | 0 | $C$ |
| 4 | I scan attachments for viruses before downloading or opening them | 3.51 | 1.43 | 0.03 | 16 | |
| 5 | I carefully handle suspicious emails even if the sender address appears trustworthy | 4.43 | 0.82 | 0.01 | - | $\mu$,$\sigma$ |
| 6 | I encrypt my email contents when sending sensitive information (e.g., banking and health information or social security number) | 2.52 | 1.56 | 0.14 | 13 | |
| 7 | I verify that my anti-virus software is up-to-date | 3.77 | 1.30 | 0.03 | 14 | |
| 8 | I turn on automatic updates for devices and applications upon installation | 3.56 | 1.22 | 0.00 | 1 | |
| 9 | When I am prompted about a device or software update, I immediately install it | 3.30 | 1.13 | 0.00 | 1 | |
| 10 | I install anti-virus software when setting up my devices | 3.91 | 1.31 | 0.02 | 6 | |
| 11 | I use browsers that protect against phishing | 4.00 | 1.21 | 0.04 | 14 | $F$ |
| 12 | I am careful about following email links that ask me to provide sensitive information | 4.71 | 0.66 | 0.00 | - | $\mu$,$\sigma$ |
| 13 | I verify links (e.g., in the URL bar or by mouseover) to ensure that I am accessing intended websites | 3.83 | 1.19 | 0.02 | 16 | |
| 14 | I avoid clicking on ads while browsing the Internet | 4.32 | 0.88 | 0.00 | - | $\mu$,$\sigma$ |
| 15 | I abstain from entering sensitive information on websites not using HTTPS | 4.07 | 1.20 | 0.03 | - | $\mu$ |
| 16 | I validate the digital certificates on the websites I visit | 2.64 | 1.38 | 0.09 | 16 | |
| 17 | I review the validity of my root certificates | 2.12 | 1.40 | 0.21 | 7 | |
| 18 | I avoid downloading and installing untrusted programs I may not be looking for | 4.69 | 0.56 | 0.00 | - | $\mu$,$\sigma$ |
| 19 | I validate the digital signatures files before opening them | 2.68 | 1.46 | 0.15 | 17 | |
| 20 | I check the extensions (e.g., .exe, .pdf) of files I download | 3.93 | 1.25 | 0.02 | 11 | |
| 21 | I turn on download notifications in my browsers | 3.69 | 1.45 | 0.05 | 4 | |
| 22 | I physically destroy drives I am done using and wish to erase | 2.60 | 1.68 | 0.13 | 5 | |
| 23 | I encrypt my devices' disks to keep my data confidential | 2.38 | 1.48 | 0.12 | 13 | |
| 24 | I avoid storing data that I do not need | 3.71 | 1.12 | 0.01 | 2 | $F$ |
| 25 | I back up the data on my devices | 3.68 | 1.18 | 0.01 | 8 | $F$ |
| 26 | I securely lock my devices with passcodes or biometrics (e.g., touch ID) | 4.21 | 1.18 | 0.02 | - | $\mu$ |
| 27 | I lock my computer when I am away from it | 3.75 | 1.43 | 0.02 | 3 | $F$ |
| 28 | I avoid accessing online banking on public computers | 4.48 | 1.13 | 0.04 | - | $\mu$ |
| 29 | I select the strictest security settings (e.g., app permissions or browser options) that are practical | 3.68 | 1.08 | 0.01 | 16 | $F$ |
| 30 | I am wary of popups and requests, even from known sources | 4.39 | 0.85 | 0.01 | - | $\mu$,$\sigma$ |
| 31 | I suspect offers that seem too good to be true | 4.67 | 0.67 | 0.00 | - | $\mu$,$\sigma$ |
| 32 | I safely store my private key for email encryption | 2.60 | 1.68 | 0.27 | 9 | |
| 33 | I report account breaches or losses to the appropriate people | 3.45 | 1.57 | 0.12 | 8 | $F$ |
| 34 | I use a password to protect my Wi-Fi network | 4.80 | 0.69 | 0.02 | - | $\mu$,$\sigma$ |
| 35 | I avoid using open Wi-Fi networks for sensitive tasks (business, banking, shopping, etc.) | 4.09 | 1.16 | 0.01 | - | $\mu$ |
| 36 | I select hard-to-guess passwords (with multiple character types, without dictionary words, etc.) | 3.99 | 1.11 | 0.00 | 11 | |
| 37 | I select different passwords for different accounts and devices | 3.93 | 1.09 | 0.00 | 3 | |
| 38 | If accounts are compromised, I change their passwords and security questions | 4.57 | 0.83 | 0.02 | - | $\mu$,$\sigma$ |
| 39 | I avoid storing passwords in ways that make them accessible to others (e.g., writing them down on notes or in files) | 4.05 | 1.22 | 0.00 | - | $\mu$ |
| 40 | I use a password manager to create and store passwords | 2.95 | 1.60 | 0.02 | 0 | $C$ |
| 41 | When possible, I use two- or multi-factor authentication | 3.83 | 1.11 | 0.01 | 14 | |
| 42 | When setting up new devices or joining new services, I change their default passwords | 4.35 | 1.02 | 0.01 | - | $\mu$ |
| 43 | I remove programs that I do not need or use | 3.93 | 0.92 | 0.00 | - | $\sigma$ |
| 44 | I use a virtual machine when opening suspicious files or websites | 1.75 | 1.17 | 0.24 | - | $\mu$ |
| 45 | I disable auto-run to prevent potentially malicious downloaded programs from running | 3.61 | 1.55 | 0.09 | 10 | |

Table 6: All items initially considered to construct ESBS. For each item, we report the mean ($\mu$) and standard deviation ($\sigma$) of response, the fraction of N/A responses, and the number of the items that exihibit high Spearman correlation ($\geq$0.3) with each item after removing items due to ceiling and floor effects. For removed items, we denote whether they were removed due to ceiling or floor effects ($\mu$), low standard deviation ($\sigma$), weak correlation with the other items($C$), or insignificant factor loadings ($F$).