Replication: Challenges in Using Data Logs to Validate Phishing Detection Ability Metrics

Casey Inez Canfield*, Alex Davis*, Baruch Fischhoff*,
Alain Forget**, Sarah Pearman*, Jeremy Thomas*
*Carnegie Mellon University, **Google
caseycan@gmail.com, alexdavis@cmu.edu, baruch@cmu.edu,
aforget@google.com, spearman@cmu.edu, thomasjm@cmu.edu

ABSTRACT
The Security Behavior Observatory (SBO) is a longitudinal field-study of computer security habits that provides a novel dataset for validating computer security metrics. This paper demonstrates a new strategy for validating phishing detection ability metrics by comparing performance on a phishing signal detection task with data logs found in the SBO. We report: (1) a test of the robustness of performance on the signal detection task by replicating Canfield, Fischhoff, and Davis (2016), (2) an assessment of the task's construct validity, and (3) evaluation of its predictive validity using data logs. We find that members of the SBO sample had similar signal detection ability compared to members of the previous mTurk sample and that performance on the task correlated with the Security Behavior Intentions Scale (SeBIS). However, there was no evidence of predictive validity, as the signal detection task performance was unrelated to computer security outcomes in the SBO, including the presence of malicious software, URLs, and files. We discuss the implications of these findings and the challenges of comparing behavior on structured experimental tasks to behavior in complex real-world settings.

1. INTRODUCTION
Maintaining security on a home computer requires knowing which security practices are most important [18] and implementing those practices, even when they may be inconsistent with users' mental models of computer security [3, 43, 44]. Users are expected to keep their software up to date (both individual programs and their operating system), avoid suspicious links and attachments (i.e. phishing attacks), choose secure passwords, and install security programs (e.g. anti-virus). Many struggle to understand and follow all these recommendations, despite good intentions.

Meanwhile, cyberattacks are becoming more varied and pervasive [39, 40], where about 1 in every 2,600 emails are phishing attacks (primarily targeted spear phishing attacks), resulting in losses of over $3 billion from business email compromise scams over the last three years [39]. Phishing attacks are no longer limited to email, but can occur over instant messenger, social media, or text messages [39]. Phishing is often used to introduce malware to a computer [37], resulting in prolonged risk. Although there are products to help protect users, none are perfect. For example, email providers use spam filters, browsers employ blacklists to block malicious websites, and security programs block and delete malicious files and software. In some cases, this requires user engagement, such as updating security programs (if automatic updating is not enabled). In other cases, such as browser blacklists, users have little control.

Growing concern over phishing risks is driving the need for timely, cost-effective measures of individuals' vulnerability. Such metrics might be derived from actual behavior or a dedicated test. Any metric faces three challenges: (a) it must differentiate between users' ability (e.g. to detect phishing emails and maintain software) and the technology in place to protect them (e.g. spam filters and blacklists, automatic updates); (b) it must account for the low base rate of phishing attacks; and (c) it must be able to extrapolate from the observed circumstances to those where users are faced with actual attacks. A simple test with predictive validity could guide targeted interventions if it provided useful performance measures.

Here, we demonstrate a new strategy for validating metrics, by triangulating performance on an experimental task with real-world system outcomes. The experimental task was developed by Canfield, Fischhoff, and Davis [4] (referred to as Canfield et al.). It extracts individual-level signal detection measures of phishing vulnerability and was demonstrated with an online mTurk sample [4]. We validate these measures using the *Security Behavior Observatory (SBO)*, a longitudinal field study that provides detailed data on a community sample of computer users' security habits over time [9, 10].

Signal detection theory (SDT), when applied to phishing detection, distinguishes between users' ability to tell the difference between phishing and legitimate emails (sensitivity or d') and bias toward identifying ambiguous emails as phishing or legitimate (response bias or c) [24]. SDT is more useful than other metrics, such as accuracy, because it accounts for the tradeoffs that people make between false negatives (missing phishing emails and potentially falling for an attack) and false positives (mistaking legitimate emails for phishing by deleting an important message or reducing the efficiency of email).

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12 -- 14, 2017, Santa Clara, California.

Here, we first replicate Experiment 1 from Canfield et al. with SBO participants in order to assess the robustness of their original (mTurk) results [4]. We then assess the construct validity of those performance metrics, in terms of correlations with self-reports on the Security Behavior Intentions Scale (SeBIS) [6]. Finally, we assess their predictive validity in comparisons with evidence of security vulnerabilities on their home computers.

Canfield et al.'s task uses realistic email messages to elicit users' detection ability and behavior. Their experiments found robust results across several experimental manipulations in mTurk samples [4]. As described below, the SBO sample is different in many ways, raising the question of how similar their performance will be. If it proves robust, one can then ask how strongly it is related to other computer security performance measures. Experimental measures are extracted under conditions where participants know that they are being observed, which may affect their behavior in various ways, including behaving so as to satisfy (or perhaps frustrate) perceived research goals [31, 35]. Here, we consider two such tests:

- *Construct validity* [5]: how well the SDT measures correlate with another theoretically related (and validated) measure, the Security Behavior Intentions Scale (SeBIS) [6]; and
- *Predictive validity*: how well the SDT measures predict actual behavior, tested by whether they improve the fit of logistic models for vulnerability to phishing attacks.

We find some evidence of robustness and construct validity, but not for predictive validity. We discuss the ways in which that failure reflects on the measures and on the challenges of characterizing vulnerability in real-world settings, shaped by users' behavior (regarding security and other matters) and their computer environment (e.g. browser, OS).

2. BACKGROUND AND RELATED WORK

The probability of users experiencing negative computer security outcomes (such as viruses) reflects both their vulnerability and their exposure [26]. The former includes their ability to detect and avoid threats (e.g. identify phishing emails), as well as their engagement in risky behavior (e.g. not updating software). The latter reflects their attractiveness as targets. We now review research on (a) measuring phishing detection performance and (b) determinants of vulnerability and exposure.

2.1 Measuring Phishing Detection Performance

There are two primary strategies for measuring users' phishing detection performance: subjective and objective. Egelman and Peer developed a subjective scale of Security Behavior Intentions (SeBIS), with four subscales: device securement, password generation, proactive awareness, and updating [6]. The proactive awareness subscale, which measures attention to URLs, has special interest for phishing vulnerability. Low scores on the proactive awareness subscale have been related to impulsivity, risk-taking, and dependence (i.e. relying on other people), consistent with the phishing detection literature [6, 34, 42, 47]. In a test of validity, Egelman, Harbach & Peer found that performance on the proactive awareness scale was correlated with respondents' ability to detect a phishing website in a laboratory environment without priming (without telling them that they were being tested on that ability) [7]. The only way to determine whether it was a phishing website was to look at the URL. Although only 22 of 718 participants correctly identified the phishing website, their proactive awareness scores were significantly higher than those of the rest of the sample [7].

Objective measures assess users' actual ability to identify phishing emails, rather than relying on self-reports of how well they do. They allow varying experimental conditions, to examine the effects of situational factors (e.g. perceived consequences, habits, stress) on phishing vulnerability [4, 41]. Canfield et al. estimated SDT parameters on an individual level by asking users to identify which emails were phishing in a set of 38 in an online test [4]. However, such tests are vulnerable to experimenter demand effects, where subjects do what they think they should do, rather than what they would normally do [27, 38].

Although controlled studies are cheaper and easier to implement than field tests, it is important to validate such measures against real world behavior. Ideally, one would send emails (both legitimate and phishing) to participants to determine how well their performance in the artificial test environment reflects their normal behavior. However, this is not always possible. It can be challenging to do high-fidelity field tests (i.e. without providing feedback on performance) without putting users at risk or incurring high costs. (Examples include Jakobsson et al. [19] and Kumaraguru et al. [20]).

As an alternative to field tests, Sotirakopoulos et al. propose examining logs of user behavior [38]. The SBO is an ongoing data collection effort that collects such logs. Its wealth of data provides multiple ways to assess vulnerability and to account for other factors that might influence users' experiences of negative computer security outcomes. Demand effects are expected to be minimal given that the study software is sufficiently unobtrusive that participants often report having forgotten that they were participating in the study. In a similar observational study, few participants reported altering their behavior in an exit survey [21].

2.2 Determinants of Vulnerability and Exposure

Unsophisticated or careless users may escape harm if they seldom use their computers or avoid dangerous situations. Conversely, knowledgeable users may ward off most attacks, yet still succumb if they use their computers heavily or are valuable targets, subject to particularly effective attacks (such as spear phishing).

Research suggests that user knowledge alone cannot compensate for the increased odds of exposure to negative computer security outcomes that come with increased use. Although research on phishing susceptibility has found that individuals with higher computer literacy are less susceptible to individual phishing attacks [36, 47], more computer-literate users tend to use their computers more frequently [2], increasing their chances of exposure to attacks and negative outcomes [22]. SBO research suggests that users' engagement with security issues, as expressed in interviews, is not a good predictor of their security outcomes [10]. Lalonde-Levesque et al. also found that more technically-savvy users are more likely to be exposed to malware threats [21]. Therefore, it is critical to control for exposure when assessing the relationship between phishing detection performance and negative computer security outcomes.

Users may also experience more negative computer security outcomes because they engage in risky behavior, such as

frequently clicking on links in emails or not updating their anti-virus software. In a survey of Dutch citizens, Leukfeldt found that while the OS type was related to malware, updated anti-virus was not [22]. While anti-virus software protection against social engineering and zero-day exploits is limited, one would expect protection against spam-type attacks using known malicious software. Our analysis assesses the relevance of this variable.

3. METHOD

3.1 Decisions in Phishing Scenarios (SDT)

Canfield et al. [4] used a scenario-based approach [20, 34], in which participants reviewed emails of a fictitious persona, Kelly Harmon. Before beginning that task, participants reviewed the PhishGuru comic strip [29] to ensure that they had some knowledge of phishing and understood their task. They then saw one of two *notifications of base rate*: "Approximately half of the emails are phishing emails" or "Phishing emails are included." *Attention* was a binary (0,1) measure, where 1 described participants who correctly answered 3 questions: "Where does Kelly Harmon work?", "What is a phishing email?", and an email that said, "If you are reading this, please answer that this is a phishing email."

Participants evaluated 38 email messages, half of which were phishing (adapted from public archives), in a random order. The base rate of phishing emails (50%) was much higher than in everyday settings (<1%) [39] in order to collect enough judgments without overburdening participants. We used the same stimuli as Canfield et al. [4] (available online at https://osf.io/7bx3n/). They ranged in difficulty from obvious phishing messages with typos to more sophisticated spear phishing attacks. For each email, participants answered the following questions:

1. *detection*: "Is this a phishing email?" (Yes/No);
2. *behavior*: "What would you do if you received this email?", with multiple-choice options including "click link/open attachment," "check sender," "check link," "reply," "ignore or archive it," "delete it," "report as spam," and "other" (following [36]);
3. *confidence*: "How confident are you in your answer?" (50-100%); and
4. *perceived consequences*: "If this was a phishing email and you fell for it, how bad would the consequences be?" (Likert scale: 1 = not bad at all to 5 = very bad).

We limited the replication to Experiment 1 in Canfield et al., which asked all participants to perform both the detection and behavior tasks. In Experiment 2, participants were randomly assigned to perform either the detection or behavior task. Canfield et al. found no significant differences in the SDT performance metrics between Experiment 1 and Experiment 2. Given the limited sample of SBO participants, having all participants perform both the detection and the behavior tasks maximized the precision of our parameter estimates. We also measured the time spent on the phishing information comic and median time spent on each email. Finally, we collected demographic information on gender, age, and education.

We evaluated individual performance using signal detection theory (SDT), a mathematical method for characterizing users' ability to distinguish phishing and legitimate emails (d') and their bias toward perceiving emails as phishing or legitimate (c). The SDT measures capture the trade-off between hit rates (*H*, correctly identifying emails as phishing) and false-alarm rates (*FA*, incorrectly identifying legitimate emails as phishing) using an inverse normal transformation to convert the probability to a Z-score:

$$d' = z(H) - z(FA)$$
$$c = -0.5(z(H) + z(FA))$$

As described by Canfield et al. [4], we estimated SDT parameters for the detection ($_D$, question (1) above) and behavior ($_B$, question (2) above) tasks separately. Thus, we calculated four phishing vulnerability parameters, summarized in Table 1.

Table 1. Phishing vulnerability parameters calculated using signal detection theory (SDT) for replication and validation of Canfield et al. [4].

Parameter	Definition
Detection Sensitivity (d'_D)	Ability to distinguish between phishing and legitimate emails.
Behavior Sensitivity (d'_B)	Ability to distinguish between when to click on links and when not to.
Detection Response Bias (c_D)	Bias toward identifying an email as phishing (negative c) or legitimate (positive c).
Behavior Response Bias (c_B)	Bias toward clicking on links (positive c) or not (negative c).

3.2 Security Behavior Intentions Scale (SeBIS)

As part of their SBO tasks, 84 participants completed the Security Behavior Intentions Scale (SeBIS) [6]. The SeBIS has 16 statements describing behaviors divided into four subscales: device securement, password generation, proactive awareness, and updating. Respondents rate on a Likert scale whether they *never* (1) to *always* (5) perform the stated behavior. Conceptually, the signal detection measures should be most closely related to the proactive awareness subscale, which includes five statements related to evaluating links, such as "When browsing websites, I mouseover links to see where they go, before clicking on them" and "I know what website I'm visiting based on its look and feel, rather than by looking at the URL bar" (reverse coded).

3.3 Home Computer Security Outcomes (SBO)

The Security Behavior Observatory (SBO) is an ongoing longitudinal study, gathering field data about home users' computer security habits. SBO participants agree to install the project software on their personal computers to gather data on their Internet browsing, installed applications, processes, network connections, system events, and more. This software then securely transmits the data to the researchers.

From these data, we measured three types of negative computer security outcomes: (a) visits to malicious URLs, (b) installed malware, and (c) presence of malicious files. Malicious URLs were identified using the Google Safe Browsing API [14] with participants' web browsing (i.e. Internet Explorer, Chrome, and Firefox) and network packet data. Due to technical limitations with browser extensions, we were unable to collect data from other popular browsers, such as Microsoft Edge. However, those data were observed in the network packet data, which include all HTTP traffic for each webpage, making it a much richer source than the browser data, which only record webpage URLs. The average webpage has approximately 100 HTTP requests for the

HTML, CSS, images, ads, multimedia, JavaScript, Flash and other files that form a single webpage [17].

We identified malware with ShouldIRemoveIt.com, which is designed to help users remove unwanted applications from their computer. We identified malicious files with VirusTotal.com, a subsidiary of Google that aggregates anti-virus scanners. For flagging software or files as malicious, we used a threshold of at least 2 scanners for ShouldIRemoveIt.com and at least 2% of scanners for VirusTotal.com. Using greater scanner agreement did not significantly change the results. Malicious files were identified across the entire computer, while malware was limited to installed applications. We assessed each outcome as a binary variable (where 1 indicates that the outcome was observed at least once and 0 indicates that the outcome was not observed), rather than a continuous one (i.e. number of negative outcomes) due to the high number of participants who had no negative outcomes (i.e. had never visited a malicious website or had no malware) and the potential unreliability of count data [23]. Participants varied in how much they used their computers, which as described above is related to the observation of negative outcomes.

We constructed logistic regression models for each outcome following the logistic model construction strategy outlined by Hosmer et al. [16] for identifying potential predictors, defined as those with statistically significant univariate correlations with the outcomes. These potential predictors are described in the next two subsections. To avoid bias and maintain transparency, we preregistered the logistic regression models at the Open Science Framework (https://osf.io/jkhbv/) before combining the SBO and SDT experiment data [25, 28]. The analysis reported here differs from the proposed analysis in the preregistration due to our acquiring more SBO data. We also improved the analysis by: (a) eliminating repetitive measures (e.g. counts of social media domains), (b) implementing an automated process for identifying malware, rather than relying on manually coded items, and (c) adding malicious files as an outcome variable.

3.3.1 Browsing exposure and risky behavior

We identified 3 variables to describe browsing exposure. Each was calculated separately for the browser and network packet data. They were (a) *total URLs/day*, (b) *unique URLs/day*, and (c) *domains/day*. Each daily count was only for days that data were received from the participant's machine.

We measured risky behavior in terms of counts of *clicked email links/day*. We expected users who clicked on more links in emails to be more likely to visit malicious URLs. We assessed this activity in 2 ways: (a) URL tracking, for URLs that include "mail" or "email" after =, &, or ? (excluding email domains), and (b) source data, where the source URL was an email domain and the destination was not. The source data did not describe whether links were clicked from an email software client, such as Microsoft Office Outlook. For the network packet data, we could only use the URL tracking method (a), because source data were unavailable.

3.3.2 Software exposure and risky behavior

We measured software exposure as a count of *total software*, excluding updates, installers, and language packages.

We sought to measure risky behavior with three variables: *delayed software updates*, *days since Windows update*, and *third-party security software* (e.g. anti-virus, anti-malware). Delayed software updates on SBO participants' computers is a count, ranging from 0 to 6, of the number of outdated popular software including Adobe Flash, Adobe Reader, Java, Internet Explorer, Chrome, and Firefox. A program was considered outdated if the participant's computer had not updated to the latest version a week after it was released. Days since Windows update is the number of days since a Windows update was most recently installed. Thus, a low number suggests that the user has updated their Windows OS more recently. This measure does not capture why users waited to install updates (e.g. whether they actively delayed updates or did not see prompts).

For third-party security software, we assigned a binary variable where 1 indicated that it was installed and error-free (see below) and 0 indicated errors or no software. Security software was considered error-free if it was in use for over 7 days, updating without errors, and scanning. In some cases, it was impossible to know if a security program met all these criteria because either it did not log the data or the log was not informative. In those cases, we used the available subset of these criteria. Thus, we assumed that installed security software was error-free unless there was evidence otherwise. We could examine the logs for McAfee, Malwarebytes, Webroot, Avast, Norton, Kaspersky, and AVG to assess their median days in use: 172 (M = 223, SD = 238). We could not assess updating for Avast or scanning for McAfee, Avast, and AVG due to missing or uninformative logs.

3.4 Sample

SBO participants were recruited from local participant pools and are predominantly retirees and college students. For this study, we recruited participants from among those who joined the SBO between October 2015 and February 2016, asking for volunteers to participate in "an online research study about email use." In addition to their regular monthly SBO compensation, each received $20 upon completing our phishing detection experiment. Those who did not start the experiment were sent a reminder after 9 days. Those who started, but did not finish, were sent a reminder after 9 days and again after another 7 days. SBO participants received higher compensation than mTurk participants ($20 vs. $5) to encourage a high response rate, given the limited pool of SBO participants. This study was approved by the Carnegie Mellon University Internal Review Board.

3.5 Defining Successful Replication

The replicability of Canfield et al. can be assessed in terms of the methods (also referred to as reproducibility) and results [13]. Canfield et al. made their original study materials and code publicly available[1] and this paper follows suit to ensure the methods are reproducible (see Appendix). The following analysis is focused on assessing whether the results are replicable and robust to changes in the study sample.

There is an ongoing debate regarding how to measure a successful replication [1, 8, 11, 29]. For this study, we assess whether the replication was successful in four ways:

1. Comparison of effect sizes
2. Consistency of the hypothesis test results
3. Parameter space region ruled out by confidence intervals
4. Combined analysis

We (1) directly compare the point estimates or effect sizes of the SDT parameters for the original and replication study. First we qualitatively compare the point estimates, considering a

meaningful difference of a 10% change in the hit rate, or probability of detecting phishing emails as unsuccessful replication. For the SDT parameters, this is a difference of 0.3 for d' and 0.1 for c. We then use a two-sample statistical significance test of the null hypothesis that the two studies were drawn from populations with the same effect size. The limitation of this first approach is that a conclusion that the study replicated based on the failure to reject the null hypothesis depends on the statistical power of the test, and thus sample size of both studies. Lower statistical power would lead to a higher frequency of conclusions that the study replicated even in the face of large differences, and high statistical power would lead to conclusions that the study did not replicate even if the differences in effect sizes were trivial.

Our second test (2) assesses the consistency of the regression coefficients in the replication study with the null hypothesis that the regression coefficient is exactly zero. The p-value on the t-test of each regression coefficient provides this measure of consistency [45]. If the p-value is below the .05 alpha level, we conclude that the regression coefficient from the replication study is inconsistent with zero, and that the study successfully replicated. The limitation of this second approach is the opposite of the first, where lower statistical power would lead to fewer conclusions that the study successfully replicated even if the regression coefficient was large, and high statistical power would lead to more conclusions that the study successfully replicated even if the regression coefficient was small.

Third, we assess (3) the region of the parameter space ruled out by confidence intervals. In the original and replication studies, we construct 95% confidence intervals. Each interval either does or does not cover the population parameter, and if we conclude that it does include the population parameter, then we will be wrong 5% of the time (i.e. the population parameter falls outside the interval). Therefore, a successful replication would find similar conclusions about the population parameter (i.e. that the region of the parameter space outside the interval in the two studies is "similar"). We operationalize this similarity as having a non-empty union of the two intervals, or that the intervals overlap. In other words, we judge that a study replicated the first if the two studies do not rule out all of the parameter space. This approach has the same limitations as the first, of always concluding successful replication with a low sample size, and never concluding successful replication with a large sample size.

Fourth, we assess (4) a combined regression analysis. We assessed whether the replication was successful by combining the two studies into a single linear regression analysis. A successful replication is then drawing the same conclusion using the combined data as the original data. This analysis improves the power of the statistical tests due to the increased sample size achieved by combining the two samples.

When considered together, these tests provide insight into whether the replication was successful. One of the primary challenges in assessing whether a replication is successful is accounting for Type II error (i.e. incorrectly accepting the null hypothesis). In the context of replication, this is the probability of incorrectly finding that the replication is successful, when in truth it is not. In this study, the sample size of the replication is constrained by the existing SBO participant pool, which limited our ability to perform a higher-powered test and increases the chance of Type II errors. To account for this, we interpret a failure to reject the null hypothesis (i.e. finding that there is no difference in effect size or hypothesis test result) as a lack of evidence of a difference, rather than evidence that there is no difference. Similarly, confidence intervals tend to be larger when the sample size and statistical power are lower, increasing the likelihood that our replication meets our definition of success. Therefore, it is critical to not over-interpret these results. Rather, this is a first attempt to use data logs for validation. As more data is collected, the strength of replication studies using this approach will improve.

3.6 Analysis

In the analysis that follows, we first reproduce the phishing detection experiment by Canfield et al. [4] to assess whether SBO participants perform differently than Amazon Mechanical Turk [32] participants (*mTurk*). We assess differences between the samples using t-tests (t), Chi-squared tests (χ^2), and 2-sided Mann-Whitney-Wilcoxon (W) tests where appropriate. Given the large number of statistical tests across disparate analyses, we generally use $\alpha = .01$ as a threshold for interpretation, rather than applying separate corrections to groups of tests. We replicate the estimation of the SDT parameters and the linear regression analysis to determine any differences in which factors predict performance. In the regression analysis, with 11 independent tests and $\alpha = .05$, we would expect to find at least one false positive (55% chance). Using $\alpha = .01$ reduces this chance to 11%. However, using $\alpha = .01$ is conservative for Type I errors, but not Type II errors. Therefore, we interpret significance using $\alpha = .05$ for the replication (where Type II error matters most) and $\alpha = .01$ for the remaining analysis (where Type I error matters most).

Second, we assess the experimental measures' construct validity with the Pearson correlation between the SDT parameters and a validated measure of security intentions, the Security Behavior Intentions Scale (SeBIS) [6].

Third, we assess predictive validity by whether the SDT parameters improve the fit of logistic models for predicting observations of negative computer security outcomes for SBO participants (i.e. observations of malicious URLs, files, and software). For each outcome, we construct a logistic regression model comprised of the SDT parameters and other predictors of exposure and risky behavior. This serves to test two hypotheses. We expect users who are more susceptible to phishing on the experimental measure to experience more negative computer security outcomes in real life. Thus, our first hypothesis is:

> *H1: Users who are more susceptible to phishing in the SDT experiment (i.e. are less able to detect and avoid threats) are more likely to visit malicious URLs and have malware and malicious files on their computer.*

We test H1 using a likelihood ratio test, which compares goodness of fit for nested logistic regression models with and without the SDT parameters. The likelihood ratio test is the most efficient test of the null hypothesis that the SDT measures do not increase the likelihood of the data given the SDT measures [15, 16]. The second hypothesis we test is:

> *H2: Users who use their computers more (i.e. have greater exposure) or engage in more risky behavior are more likely to visit malicious URLs as well as have malware and malicious files on their computer.*

We test H2 in the construction of the logistic regression models, following the procedure recommended by Hosmer et al. [16].

4. RESULTS

4.1 Sample

We recruited 132 SBO participants to participate in the phishing detection experiment. Of those, 121 started the survey and 98 finished (= 74% response rate). We excluded 5 participants who sent the SBO less than 7 days of data. The final sample (see SBO Sample in Table 2) represents 44% (= 93/213) of all the SBO participants at that time (All SBO in Table 2). As shown in Table 2, the SBO sample was older, $t(121) = 4.52$, $p < .001$, Cohen's $d = 0.69$, and had a higher proportion of college-educated individuals, $\chi^2(1) = 6.83$, $p = .009$, $\phi = 0.17$, than did the mTurk sample in Canfield et al. [4].[1] There was no difference in gender, $\chi^2(1) = 0.05$, $p = .823$, $\phi = 0.01$. Within the SBO sample, older participants tended to be better educated, in part because some of the younger participants were in college (thus had not finished their educations), $r(92) = 0.37$, $p < .001$. Our SBO sample resembled the wider SBO population on these variables (Table 2).

Table 2. Comparison of mTurk and SBO demographics. The mTurk sample is from Canfield et al. [4].

Variable	mTurk	SBO Sample	All SBO
Female	58%	60%	61%
Bachelors+	45%	63%	58%
Age	32 [19, 59]	41 [19, 81]	46 [19, 87]
N	152	93	213

4.2 Comparison of Experimental Results (Replication)

There was little difference between how much attention the SBO and mTurk participants paid to instructions. Of the 93 SBO participants, 16 failed at least 1 of the 3 attention checks. Users who failed the attention checks were not excluded from the sample, but attention was included as a variable in the regression analysis in order to increase statistical power [30]. There were no significant differences in performance on the attention checks, 17% failed for SBO vs. 10% failed for mTurk, $\chi^2(1) = 2.18$, $p = .14$, $\phi = 0.09$. The median time spent on the introductory phishing information was slightly higher for the mTurk participants, *SBO* = 0.74 minutes ($M = 1.16$, $SD = 1.79$) vs. *mTurk* = 0.95 minutes ($M = 3.17$, $SD = 11.51$), $W = 5018$, $Z = 2.25$, $p = .02$, $r = 0.14$.

However, SBO participants, particularly the older ones, spent more time on the individual email stimuli. The median time to complete the experiment was 47 minutes, including breaks ($M = 59$ min, $SD = 49$ min). This estimate excludes seven outliers, participants who appeared to stop working and leave the experiment open on their browser for 19 hours to almost 2 weeks. SBO participants spent more time per email, *SBO* = 0.94 minutes ($M = 1.13$, $SD = 0.72$) vs. *mTurk* = 0.48 minutes ($M = 0.53$, $SD = 0.24$), $W = 11850$, $Z = 8.88$, $p < .001$, $r = 0.57$ in a Mann-Whitney-Wilcoxon test. Within the SBO sample, older participants spent more time per email, $r(92) = 0.46$, $p < .001$.

First, we assess whether the results of the SDT parameter estimation replicate. Since these are point estimates, there are no hypothesis tests to replicate. There was no evidence of significant differences between the mTurk and SBO samples on any SDT parameters, for either the detection or the behavior task, $p > .05$. However, the point estimates differ by 0.12 for detection c, which exceeds our meaningful difference threshold. When comparing the confidence intervals, the replicated point estimate is within the original study's confidence interval for d' and behavior c. For detection c, the replicated point estimate is outside of the original confidence interval, but the confidence intervals still overlap. In general, there is no evidence that the SDT estimates differ between the studies, although the evidence is weakest for detection c. Table 3 shows the mean statistics for the SDT parameters and accuracy (for comparison). Figure 1 shows the distribution of d' and c for each task and sample. There was no evidence of learning over the course of the experiment, as d' and c were equal when calculated separately for the first and second half of the emails. This suggests that the performance parameters estimated in Canfield et al. [4] are not unique to mTurk and can be generalized to the SBO population, which was an older, potentially less tech-savvy group.

We also replicated the regression analysis from Canfield et al. [4] to determine whether there were any differences in the factors that predicted phishing vulnerability for the two samples. Tables 4 and 5 show the results for both samples to compare the results of the hypothesis tests. Figure 2 compares the 95% confidence intervals. In general, the SBO sample's coefficients had larger confidence intervals, due to the lower sample size, but overlap the mTurk coefficients, suggesting no statistically significant differences. The results were largely the same, except for the following three differences.

First, unlike Canfield et al.'s mTurk sample, confidence was not a significant predictor of response bias (c) for the SBO sample. We found no systematic differences in mean confidence between the two samples, $M = 0.86$ ($SD = 0.08$) for SBO and mTurk, $t(181) = 0.04$, $p = .97$, Cohen's $d = 0.01$. Second, age and education are predictors of c in the SBO sample, but were not in the mTurk sample, perhaps due to the higher variance of age and education in the SBO sample. Older participants seemed biased toward identifying emails as phishing (i.e. lower detection c). College-educated participants seemed biased toward identifying emails as legitimate (higher detection c). Third, attention and median time per email were not significant predictors for the SBO sample, perhaps due to reduced variance, as SBO participants were more likely to fail the attention checks and spent more time per email.

As also reported in Tables 4 and 5, the combined analysis is largely consistent with the original Canfield et al. experiment for sensitivity, but there are differences for response bias. Higher attention and higher average confidence predict higher detection sensitivity, consistent with the original Canfield et al. ($p < .01$). None of the predictors are significant for behavior sensitivity, consistent with the original Canfield et al. ($p < .01$). Higher average confidence, lower perceived consequences, and younger individuals tended to have a higher detection response bias, which differs from the original Canfield et al. study ($p < .01$). In the separate analysis, age is significant for the SBO sample but not the mTurk sample and average confidence is significant for the mTurk sample but not the SBO sample. Higher average confidence and lower perceived consequences are associated with a higher behavior response bias, which differs from the original Canfield et al. study ($p < .01$). In the separate analysis, the median time spent per email is significant for the mTurk sample and none of the predictors are significant for the SBO sample.

Table 3. SDT phishing vulnerability parameter estimates for mTurk [4] and SBO samples.

	Detection Task (Yes/No)			Behavior Task (multiple choice)			
	mTurk M (SD) [CI]	**SBO M (SD) [CI]**		**mTurk M (SD) [CI]**	**SBO M (SD) [CI]**		**Typical Range**
Sensitivity (d')	0.96 (0.64) [0.86, 1.06]	0.96 (0.66) [0.83, 1.10]	$t(191) = 0.01$, $p = .99$, $d = 0$	0.39 (0.50) [0.31, 0.47]	0.42 (0.52) [0.32, 0.53]	$t(190) = 0.41$, $p = .69$, $d = 0.05$	0 to 4
Response bias (c)	0.32 (0.46) [0.24, 0.39]	0.20 (0.51) [0.10, 0.30]	$t(178) = -1.78$, $p = .08$, $d = 0.24$	-0.54 (0.66) [-0.64, -0.43]	-0.62 (0.57) [-0.74, -0.51]	$t(216) = -1.07$, $p = .29$, $d = 0.14$	-2 to 2
Accuracy	0.67 (0.11) [0.65, 0.69]	0.67 (0.11) [0.65, 0.69]	$t(193) = 0.03$, $p = 0.98$, $d = 0$	0.56 (0.08) [0.55, 0.57]	0.57 (0.09) [0.55, 0.59]	$t(179) = 0.99$, $p = .32$, $d = 0.13$	0 to 1

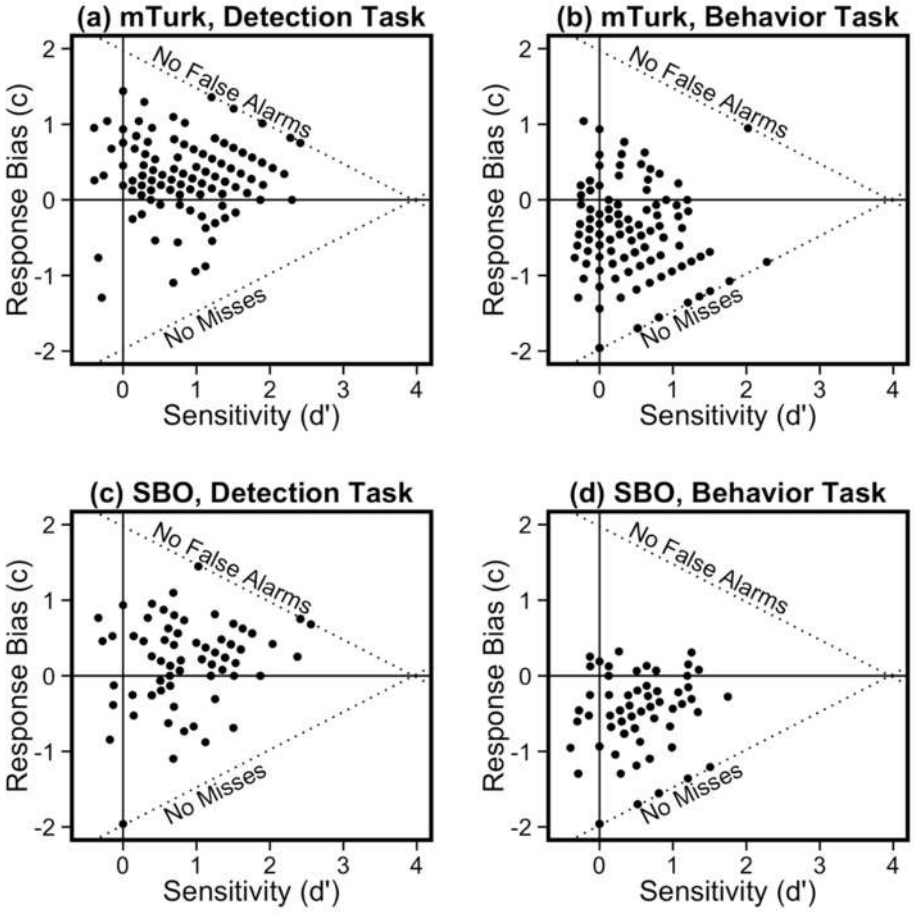

Figure 1. Plot of d' vs. c for each task and sample. The parameter estimates are bounded by the dotted lines, which represent extreme performance (no false alarms or no misses). There were no significant differences in performance between the mTurk (a, b) [4] and SBO (c, d) samples.

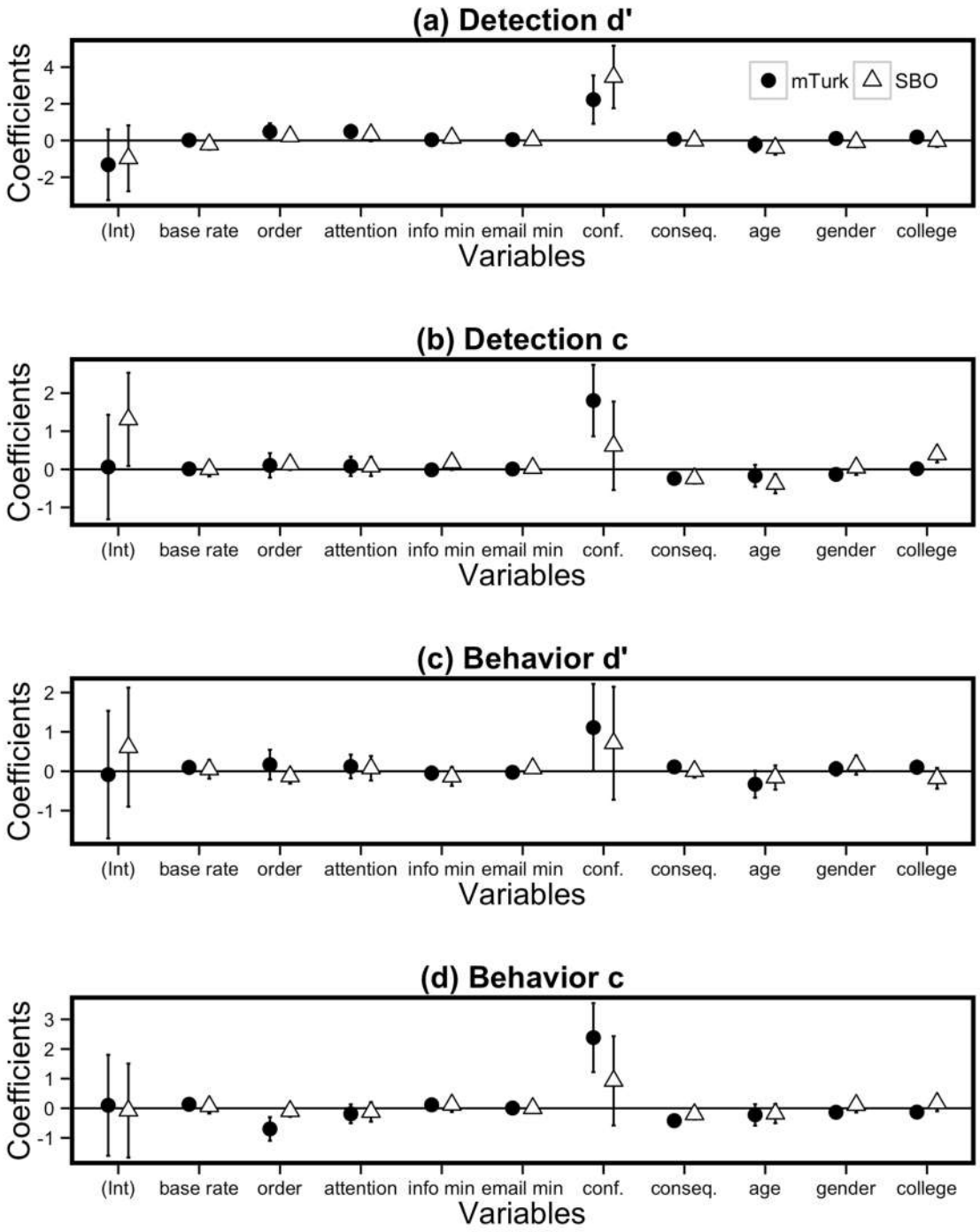

Figure 2. Comparison of regression coefficients with 95% confidence intervals (CI) for (a) detection d', (b) detection c, (c) behavior d', and (d) behavior c.

Table 4. Comparison of linear regression analysis of detection and behavior sensitivity (d') for mTurk [4] and community (SBO) samples. The asterisks indicate statistical significance, where * is p < .05, ** is p < .01, and * is p < .001.**

	Detection Sensitivity (d'_D)			Behavior Sensitivity (d'_B)		
	mTurk B (SE)	SBO B (SE)	Combined B (SE)	mTurk B (SE)	SBO B (SE)	Combined B (SE)
Intercept	-1.32 (0.98)	-0.97 (0.92)	-0.96 (0.64)	-0.09 (0.83)	0.61 (0.77)	0.19 (0.54)
Sample (SBO = 1)			-0.04 (0.10)			0.15 (0.08)
Knowledge of base rate	0.02 (0.10)	-0.22 (0.14)	-0.05 (0.08)	0.10 (0.08)	0.05 (0.12)	0.08 (0.07)
Task order (detection = 1)	0.04 (0.10)	0.15 (0.14)	0.09 (0.08)	-0.05 (0.09)	-0.14 (0.12)	-0.06 (0.07)
Attention (pass = 1)	0.49 (0.18)**	0.33 (0.19)	0.40 (0.13)**	0.12 (0.15)	0.07 (0.16)	0.10 (0.11)
log(Phish info time)	0.05 (0.04)	0.02 (0.07)	0.05 (0.03)	-0.03 (0.03)	0.07 (0.06)	0 (0.03)
Median time/email	0.48 (0.23)*	0.23 (0.11)*	0.21 (0.09)*	0.17 (0.19)	-0.13 (0.09)	-0.08 (0.08)
Average confidence	2.23 (0.67)**	3.46 (0.87)***	2.64 (0.51)***	1.11 (0.57)	0.71 (0.73)	0.84 (0.43)
Average perceived consequences	0.08 (0.08)	0 (0.10)	0.07 (0.06)	0.11 (0.06)	0 (0.08)	0.08 (0.05)
log(Age)	-0.22 (0.21)	-0.40 (0.19)*	-0.33 (0.13)*	-0.33 (0.17)	-0.16 (0.16)	-0.26 (0.11)*
Gender (male = 1)	0.11 (0.10)	-0.09 (0.15)	0.04 (0.08)	0.06 (0.09)	0.15 (0.12)	0.10 (0.07)
College (college degree = 1)	0.19 (0.10)	-0.03 (0.16)	0.11 (0.08)	0.10 (0.09)	-0.18 (0.13)	0.02 (0.07)
N	142	84	227	142	84	227
Adjusted R^2	0.16	0.14	0.15	0.05	0.05	0.05
F	3.71***	2.37*	4.63***	1.68	1.40	2.16*

Table 5. Comparison of linear regression analysis of detection and behavior response bias (c) for mTurk [4] and community samples. The asterisks indicate statistical significance, where * is p < .05, ** is p < .01, and * is p < .001.**

	Detection Response Bias (c_D)			Behavior Response Bias (c_B)		
	mTurk B (SE)	SBO B (SE)	Combined B (SE)	mTurk B (SE)	SBO B (SE)	Combined B (SE)
Intercept	0.06 (0.70)	1.31 (0.62)*	0.81 (0.46)	0.10 (0.87)	-0.08 (0.81)	-0.14 (0.58)
Sample (SBO=1)			-0.12 (0.07)			0.13 (0.09)
Knowledge of base rate	0.01 (0.07)	0 (0.10)	-0.01 (0.06)	0.13 (0.09)	0.07 (0.13)	0.10 (0.07)
Task order (detection=1)	-0.01 (0.07)	0.18 (0.10)	0.01 (0.06)	0.11 (0.09)	0.12 (0.13)	0.08 (0.07)
Attention (pass = 1)	0.08 (0.13)	0.07 (0.13)	0.07 (0.09)	-0.19 (0.16)	-0.13 (0.17)	-0.13 (0.12)
log(Phish info time)	0.01 (0.03)	0.04 (0.05)	0.01 (0.02)	0.01 (0.04)	0 (0.06)	0 (0.03)
Median time/email	0.10 (0.16)	0.13 (0.08)	0.14 (0.06)*	-0.70 (0.20)***	-0.10 (0.10)	-0.17 (0.08)*
Average confidence	1.81 (0.48)***	0.62 (0.59)	1.30 (0.36)***	2.38 (0.59)***	0.93 (0.77)	1.92 (0.47)***
Avg perceived consequences	-0.24 (0.05)***	-0.24 (0.07)***	-0.26 (0.04)***	-0.42 (0.07)***	-0.20 (0.09)*	-0.36 (0.05)***
log(Age)	-0.17 (0.15)	-0.38 (0.13)**	-0.27 (0.09)**	-0.22 (0.18)	-0.18 (0.16)	-0.21 (0.12)
Gender (male=1)	-0.13 (0.07)	0.05 (0.10)	-0.06 (0.06)	-0.14 (0.09)	0.11 (0.13)	-0.05 (0.08)
College (college degree=1)	0.02 (0.07)	0.39 (0.11)***	0.12 (0.06)*	-0.13 (0.09)	0.18 (0.14)	-0.02 (0.08)
N	142	84	227	142	84	227
Adjusted R^2	0.18	0.27	0.21	0.39	0.07	0.25
F	4.16***	4.12***	6.44***	9.85***	1.63	7.81***

4.3 Construct Validity

We assessed construct validity as the correlation between the SDT parameters and the proactive awareness subscale of the SeBIS. One of the four SDT parameters, behavior c (i.e. how suspicious a link must be before the participant chooses not to click on it), was correlated with the SeBIS proactive awareness subscale, $r(83) = -0.29$, $p = .008$. None of the other SDT parameters had a correlation greater than 0.20. Thus, participants who reported looking at the URL before clicking on links (in the SeBIS) were also more cautious in the experimental task (behavior c).

4.4 Predictive Validity

For simplicity's sake, we only report tests of predictive validity for the behavior task, as results for the detection task were similar. Below, we report our analyses separately for each of the four SBO computer security negative outcomes.

4.4.1 Malicious URLs in Browser Data

Browser data were available for 86 of the 93 SBO users. Most used Internet Explorer (66/86 = 77%), followed by Chrome (29/86 = 34%) and Firefox (12/86 = 14%). Some participants used multiple browsers, so the percentages do not sum to 100%. In total, 9 participants (10%) had visited a malicious URL: 2 Internet Explorer users (2/66 = 3%), 4 Chrome users (4/29 = 14%), and 3 Firefox users (3/12 = 25%).

Table 6 shows our univariate analyses [16] for the relationship between each potential predictor and whether users had visited a malicious URL. Among these potential covariates, only domains/day was related to whether participants had visited malicious URLs. Therefore, it was included in the regression model, using a log transformation to normalize the observations.

Users who visited more domains were more likely to have visited a malicious URL. Table 8 shows the regression analysis for the

browser data. Log(domains/day) was the only significant predictor. As seen in the likelihood ratio test (reported in the last row of Table 8), users' SDT parameter estimates did not improve the model fit. This indicates that there was no evidence that ability to identify phishing emails in the experiment (as represented by the SDT parameters) was related to whether participants had visited a malicious URL in the browser data.

4.4.2 Malicious URLs in Network Packet Data

We also assessed visits to malicious URLs in the network packet data. There was much more network packet data than browser data (Table 6), since a single webpage is assembled from many network packets [17]. For 31 of 93 SBO users (33%), the network packet data indicated that they had visited a malicious URL. Univariate analysis [16] found that total URLs/day, unique URLs/day, and domains/day were related to having visited a malicious URL at least once. We then computed a factor analysis, which revealed that these covariates loaded on one factor, alpha = 0.79. We called this factor *browsing intensity* and used a log transformation to normalize it. We then used that factor score in the regression model and likelihood ratio test reported in Table 8.

The regression analysis shows that users with higher browsing intensity were more likely to have visited a malicious URL in the network packet data. In addition, there was an effect for gender, whereby men were more likely to have visited malicious URLs. This finding emerges after normalizing for exposure (in the regression analysis) and observing no correlation between gender and exposure, $r(90) = .06$, $p = .57$. This suggests that men were either more likely to visit malicious URLs in their browsing or worse at detecting malicious URLs in this sample. More research is needed to understand this result. In the likelihood ratio test, users' SDT parameter estimates did not improve the model fit. Thus, there was no evidence to suggest that performance on the SDT experiment was related to whether participants had visited a malicious URL in the network packet data.

Table 6. Descriptive statistics and factor analysis for the browser and network packet sensor predictors.

	Browser		Network Packet		
	Median	M (SD)	Median	M (SD)	Loading
Days	40	67 (76)	70	85 (63)	NA
Total URLs	22	56 (90)	1,500	2,600 (3,600)	0.73
Unique URLs	9	23 (32)	670	990 (1,000)	1
Domains	5	5.7 (4.4)	42	52 (37)	0.60
% of Total Variance					63%
Cronbach's Alpha					0.80

4.4.3 Malware

Most users had the Windows 10 operating system (53/92 = 58%), followed by Windows 8 (22/92 = 24%), Windows 7 (14/92 = 15%), and Windows Vista (3/92 = 3%). 43 of the 92 (47%) users with installed software data had malware. For each operating system, approximately half of the users had malware.

Table 7 shows descriptive statistics for viable software covariates. Univariate analysis [16] revealed that total software and delayed software updates were related to malware. However, the factor analysis found that these variables were only weakly related. When included in the regression model separately, delayed software updates were not a significant predictor, so it was removed from the model. Total software was normalized using a log transformation.

Users who installed more software were more likely to have malware on their machine. As shown in Table 8, this variable predicted malware. Again, the SDT parameter estimates did not improve the model fit. Thus, there was no evidence that performance on the SDT experiment was related to observations of malware on a participant's computer.

4.4.4 Malicious Files

Most users (84/93 = 90%) had malicious files on their computer. In the regression model, we used the same predictors as in the malware model, reported in Table 7.

The regression analysis (Table 8) shows that users who had installed more software were significantly more likely to have malicious files on their computer. The SDT parameter estimates did not improve the model fit. Thus, there was no evidence that performance on the SDT experiment was related to observations of malicious files on a participant's computer.

Table 7. Descriptive statistics and factor analysis for the software predictors.

	Median	M (SD)	Loading
Total Software	244	342 (316)	0.44
Delayed Software Updates	2	2 (1)	0.44
% of Total Variance			20%
Cronbach's Alpha			0.33

5. DISCUSSION

In this study, we reproduced Experiment 1 from Canfield et al. [4] in a community sample (SBO). We assessed replicability in terms of the effect sizes, results of the hypothesis tests, confidence intervals, and combined analysis. In general, we found similar distributions of the SDT performance measures as in the mTurk sample, suggesting that there was no evidence of differences in performance between the two samples. However, although the performance of the two samples replicated (as defined in Section 3.5), the regression analysis differed slightly, reflecting the differences between the samples in terms of age and education. This analysis suggests that a higher-powered study with a diverse sample is needed to assess demographic effects. However, the findings about confidence and perceived consequences are fairly consistent, suggesting that they may be useful parameters for future behavioral interventions and predictive metrics.

We found some evidence of construct validity for the experimental behavior task, consistent with it measuring what it claimed. Participants with a greater response bias on the behavior task (c_B), or tendency to treat uncertain emails as phishing, had higher scores on the SeBIS proactive awareness subscale, which elicits self-reports of attention to URLs. This suggests that participants were acting on their computer security intentions in the SDT experiment. The other SDT parameters were not correlated with SeBIS. This suggests that ability (d') is not related

Table 8. Logistic regression models and likelihood ratio test (LRT) for each outcome. The LRT compares the full models shown above with the same models excluding the 2 SDT parameters. The asterisks indicate statistical significance, where * is p < .05, ** is p < .01, and * is p < .001.**

	Malicious URLs (browser)	Malicious URLs (network packet)	Malware	Malicious Files
(Int)	-6.43 (2.14)**	-10.53 (2.83)***	-5.93 (1.71)***	-6.65 (3.71)
Behavior Sensitivity (d'_B)	-0.06 (0.89)	-0.33 (0.55)	-0.09 (0.46)	-1.59 (1.04)
Behavior Response Bias (c_B)	-0.80 (0.74)	0.11 (0.50)	-0.06 (0.44)	-0.90 (1.22)
log(Domains/day)	1.93 (0.77)*			
log(Browsing Intensity)		1.39 (0.38)***		
log(Total Software)			0.99 (0.31)**	2.58 (0.87)**
Age	0.01 (0.03)	-0.03 (0.02)	0 (0.01)	-0.05 (0.03)
Male	0.90 (0.81)	1.47 (0.55)**	0.07 (0.48)	-0.64 (0.94)
College	-0.89 (0.95)	0.16 (0.61)	0.56 (0.53)	-1.29 (1.29)
LRT	$\chi^2(2) = 1.29$, $p = 0.5$	$\chi^2(2) = 0.41$, $p = 0.8$	$\chi^2(2) = 0.06$, $p = 1.0$	$\chi^2(2) = 4.12$, $p = 0.13$

to security intentions. The response bias (c) for the detection task measures participants' tendency to identify emails as phishing or legitimate. Although this could have been related to security intentions, the behavior task better matched the SeBIS scale due to the higher consequences associated with behavior.

We found no evidence of predictive validity for the SDT parameters for any of the four computer security outcomes in the SBO data: browser visits to malicious URLs and network packet data, malware, and malicious files. Thus, we reject H1. However, those four measures were robust enough to be predicted by other observation-based measures, as hypothesized by H2. SBO participants who used their computers more frequently were more likely to have experienced a negative computer security outcome.

We offer four possible reasons why the ability to identify suspicious messages in the laboratory task did not predict the ability to identify similar suspicious messages in the real world:

1. the experimental task does not evoke true phishing behavior,
2. the experimental task evokes true behavior in an environment different from SBO users' (i.e. lack of ecological validity),
3. the SBO measures are confounded by other aspects of users' complex real-world experience, or
4. the SBO data are too noisy to reveal the underlying correlations without much larger samples.

Explanation (1), that the experiment does not evoke actual behavior, seems unlikely, as the results of the experiment are in line with other phishing susceptibility research. For example, participants who perceived worse consequences were more cautious (negative c) [34, 42, 47]. Moreover, performance on the SDT experiment showed expected correlations with other variables, such as better performance being associated with greater security intentions (in the test of construct validity).

Explanation (2), lack of ecological validity for the experiment environment, seems more plausible. One unrepresentative feature of the experimental task is that it has a 50% base rate of phishing emails, much higher than that in everyday life [34]. That higher rate seems likely to have influenced the SDT estimates. In a SDT study of baggage screening, artificially high base rates decreased c (i.e. encouraged participants to be more biased toward identifying items in baggage as suspicious), but had no effect on d' (i.e. people's ability to differentiate between suspicious and benign items in baggage) [46]. Analogous behavior here would have been a greater propensity to treat messages as phishing in the experiment than in life. A second feature of the experimental task is explicitly asking participants to evaluate each email for phishing, thereby priming them to detect attacks. Research by Parsons et al. [33] suggests that explicitly mentioning phishing artificially increases d' but has no effect on c. Together, these studies suggest that our estimates of performance are better than what would be expected in real life. However, there is no obvious reason why these differences should affect users' relative performance. Thus, we would expect users who are good at detecting phishing to perform better on the experiment than users who are bad at detecting phishing. As a result, the correlations across measures should be preserved. In other words, we would not expect users who are bad at detecting phishing in real life to be better at it in an experiment, compared to users who are good at detecting phishing in real life.

Explanation (3), that the complexity of real-world environments (for SBO participants, among others) complicates the relationship between individuals' general propensities (which the SDT metrics attempt to measure) and their actual experiences, is also compelling. As seen here, negative experiences (in the sense of visiting malicious URLs and having malicious files) are strongly related to the amount of exposure (as measured by browsing intensity and total software). Perhaps individuals' exposure to threats overwhelms their ability (d') or propensity (c) to avoid them. Alternatively, the ability to detect phishing emails may not translate to users' ability to avoid attack vectors in general. Thus, the effect of avoiding threats from phishing is washed out by all the other attacks that lead to malware and malicious files on users' computers. Participants' rate of negative experiences may also be related to their systems' protections and their attractiveness as targets for attackers. Systems' vulnerability is partially determined by users (e.g. abilities, knowledge) and partly by others (e.g. browser blacklists, security software). Unfortunately, even with the rich SBO data set, we lacked the complete picture needed to sort out these relationships. The SBO collects data on browser warnings, but there were very few observations. Examining browser warnings would allow observation of the URLs that users attempted to visit, rather than being limited to the successful ones that were not blocked by browser blacklists. In addition, as described in the Methods section, we were unable to measure

detections for all security software. Some of those programs, particularly free versions, do not record logs. Others have poor documentation. On those that do provide logs, we observed few detections. Given that security software use did not predict the presence of malware or malicious files and that more malware and malicious files were observed than were detected by security software, one possible explanation is that many SBO users were unable to configure and utilize their security software effectively.

Finally, explanation (4), that the SBO measures are noisy, is to be expected for real-world observations. There were cases where data were missing (e.g. a sensor malfunctioned or was turned off) or ambiguous (e.g. multiple people using the same computer). As a partial check on one potential source of noise, we repeated the analysis after excluding computers with multiple users, but found similar results. If data problems are randomly distributed, then a larger sample might reveal underlying relationships. If they are correlated with individual or system performance, then those interdependencies will need to be understood and unraveled.

Thus, validating predictive measures of phishing vulnerability (including SDT and SeBIS) requires a much more nuanced picture than we currently have of the relationship between individuals' ability, propensities, and experienced outcomes. The predictive validity of any measure could be undermined by proper environmental safeguards or if people realize their vulnerability and restrict their behavior. Once available, a full picture of the SBO data may provide valuable guidance on these possibilities.

5.1 Limitations

This study had several notable limitations. First, it was limited to Windows users. The depth and breadth of SBO data collection requires custom software tailored to each OS. Due to resource constraints, the SBO is limited to Windows, the most common OS [9]. In the original mTurk sample [4], 84% of participants used Windows and performed similarly to other OS users.

Second, although this study evaluates the generalizability of an existing method, it leaves some aspects of generalizability open to further study. Although the mTurk and SBO samples differed in some ways (Table 2), extension to other populations would be valuable. One within-sample result bearing further attention is the observation that men were more susceptible to phishing.

Third, both the experimental task and the SBO study whether individuals visit a phishing website. That leaves open the question of when they share personal information once there. As noted, even the simpler outcome of such visits was difficult to measure in the SBO. We were limited by the data available in the Google Safe Browsing, ShouldIRemoveIt.com, and VirusTotal datasets. Thus, we missed attacks absent in these databases. In addition, we observed more negative computer security outcomes related to software (47% had malware and 90% had malicious files) than to browsing (10% in browser data and 33% in network packet data). This lower rate may reflect limits to the lists of malicious URLs, which change over time. For example, a legitimate site may be compromised and only briefly appear on the Google Safe Browsing blacklist. Finally, some SBO data were missing for technical reasons, which reduced our ability to observe negative outcomes and correlate them with other measures.

5.2 Recommendations

Given the novelty of using data logs like those collected by the SBO to validate performance tests like those collected in Canfield et al., we provide recommendations for future work:

1. To the extent possible, use behavioral outcomes that are (a) as directly related to the outcome of interest as possible and (b) rely on human ability without intervening technology. For example, measure attempts to visit malicious URLs (via browser warning data), rather than actual visits, to distinguish human ability from browser blacklist effectiveness. When possible, use security software detections of malware and malicious files to assess attempts to download malicious files. Technical constraints and lack of observations limited our ability to use these outcomes.
2. Triangulate between multiple data sources (e.g. assessing both browser and network packet data), with an understanding of their respective strengths and weaknesses. For example, there are more network packet data, but browser data better reflect the URLs that users choose to visit. Beyond the analysis presented here, it may be possible to crosscheck events such as security software scanning with observed active processes on the machine.
3. Consider the temporal sequence of events, such as how periods without security software protection affect the risk of acquiring malicious files.

6. CONCLUSION

We assessed the validity of the SDT measures proposed by Canfield et al. [4] in three ways: (a) replicating their mTurk SDT experiment with SBO participants, (b) assessing construct validity via correlation with the SeBIS proactive awareness subscale, and (c) evaluating predictive validity using negative outcomes observed in the SBO data. Our results suggest (a) that the findings from Canfield et al. [4] generalize to the SBO population and (b) the SDT measures have construct validity, given the correlation between participants' self-reported tendency to look at URLs before clicking links (in the SeBIS) and their caution in clicking links in the SDT study (behavior c). However, we found (c) no evidence of predictive validity, as the SDT measures did not predict negative computer security outcomes observed in the SBO.

One of the primary challenges for this analysis was differentiating between people's ability to protect themselves (by knowing which URLs to avoid) and technical safeguards (such as browser blacklists). Future research, addressing this complication, will offer opportunities for laboratory and observational measures to complement one another in understanding the security ecosystem.

7. ACKNOWLEDGEMENTS

We thank the SBO PIs: Alessandro Acquisti, Nicolas Christin, Lorrie Cranor, Serge Egelman, and Rahul Telang, for providing access to the SBO data. In addition, we thank Lorrie Cranor and Serge Egelman for comments on an earlier draft. We also thank Rick Wash for helpful comments. The Security Behavior Observatory was partially funded by the NSA Science of Security Lablet at Carnegie Mellon University (contract #H9823014C0140); the National Science Foundation, Grant CNS-1012763 (Nudging Users Towards Privacy); and the Hewlett Foundation, through the Center for Long-Term Cybersecurity (CLTC) at the University of California, Berkeley.

8. REFERENCES

[1] Anderson, C. J., Bahnik, S., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R. … Zuni, K. (2016). "Response to Comment on 'Estimating the reproducibility of psychological science,'" *Science* 351(6277), 1037-1039. DOI: 10.1126/science.aad9163

[2] Appel, M. (2012). "Are heavy users of computer games and social media more computer literate?" *Computers & Education* 59(4), 1339–1349. DOI: http://doi.org/10.1016/j.compedu.2012.06.004

[3] Camp, L. J. (2009). "Mental models of privacy and security," *IEEE Technology and Society Magazine* 28(3), 37–46. DOI: http://doi.org/10.1109/MTS.2009.934142

[4] Canfield, C. I., Fischhoff, B., & Davis, A. (2016). "Quantifying Phishing Susceptibility for Detection and Behavior Decisions," *Human Factors* 58(8), 1158-1172. DOI: http://doi.org/10.1177/0018720816665025

[5] Cronbach, L. J. & Meehl, P. E. (1955). "Construct validity in psychological tests," *Psychological Bulletin* 52(4), 281-302. http://psychclassics.yorku.ca/Cronbach/construct.htm.

[6] Egelman, S. & Peer, E. (2015). "Scaling the Security Wall," In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI '15*, 2873–2882. DOI: http://doi.org/10.1145/2702123.2702249

[7] Egelman, S., Harbach, M., & Peer, E. (2016). "Behavior Ever Follows Intention?" In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI '16*,1-5. DOI: http://doi.org/10.1145/2858036.2858265

[8] Etz, A. & Vandekerckhove, J. (2016). "A Bayesian Perspective on the Reproducibility Project: Psychology," *PLOS ONE*. https://doi.org/10.1371/journal.pone.0149794

[9] Forget, A., Komanduri, S., Acquisti, A., Christin, N., Cranor, L. F., & Telang, R. (2014). "Security Behavior Observatory: Infrastructure for long-term monitoring of client machines," Technical Report CMU-CyLab-14-009, CyLab, Carnegie Mellon University, Pittsburgh, PA. https://www.cylab.cmu.edu/files/pdfs/tech_reports/CMUCyLab14009.pdf

[10] Forget, A., Pearman, S., Thomas, J., Acquisti, A., Christin, N., Cranor, L. F., Egelman, S., Harbach, M., & Telang, R. (2016). "Do or Do Not, There Is No Try: User Engagement May Not Improve Security Outcomes," In *Proceedings of the Symposium on Usable Privacy and Security: SOUPS '16*, 97–111.

[11] Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). "Comment on 'Estimating the reproducibility of psychological science,'" *Science* 351(6277), 1037-1039. DOI: 10.1126/science.aad7243

[12] Global Stats. (2016). "Top 7 OSs," http://gs.statcounter.com/#desktop-os-ww-monthly-201610-201610-bar

[13] Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). "What does research reproducibility mean?" *Science Translational Medicine* 8(341), 1-7. DOI: 10.1126/scitranslmed.aaf5027

[14] Google. (2016). "Google Safe Browsing APIs (v4)". https://developers.google.com/safe-browsing/v4/

[15] Hauck, W. W. & Donner, A. (1977). "Wald's Test as Applied to Hypotheses in Logit Analysis," *Journal of the American Statistical Association* 72(360), 851-853.

[16] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.), Hoboken, NJ: John Wiley & Sons.

[17] HTTP Archive. (2016). "Trends." Retrieved on June 29, 2016 from http://httparchive.org/trends.php#bytesTotal&reqTotal.

[18] Ion, I., Reeder, R., & Consolvo, S. (2015). "'...no one can hack my mind': Comparing Expert and Non-Expert Security Practices," In *Proceedings of the Symposium on Usable Privacy and Security: SOUPS '15,* 327–346.

[19] Jakobsson, M. & Ratkiewicz, J. "Designing ethical phishing experiments: a study of (rot13) ronl query features," In *Proceedings of the 15th International Conference on World Wide Web: WWW '06*, 513–522.

[20] Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., & Hong, J. (2010). "Teaching Johnny not to fall for phish," *ACM Transactions on Internet Technology* 10(2), 1–31.

[21] Lalonde Levesque, F., Nsiempba, J., Fernandez, J. M., Chiasson, S. & Somayaji, A. (2013). "A clinical study of risk factors related to malware infections," In *Proceedings of the ACM SIGSAC Conference on Computer & Communications Security: CCS '13*, 97–108. DOI: http://doi.org/10.1145/2508859.2516747

[22] Leukfeldt, E. R. (2015). "Comparing victims of phishing and malware attacks," *International Journal of advanced studies in Computer Science and Engineering* 5(5), 26-32.

[23] Long, L. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*, Thousand Oaks, CA: Sage Publications.

[24] Macmillan, N. A. & Creelman, D. C. (2004). *Detection Theory: A User's Guide*, New York, NY: Psychology Press.

[25] Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., … Van Der Laan, M. (2014). "Promoting transparency in social science research," *Science* 343, 30-31.

[26] National Institute for Standards and Technology (NIST). (2012). "Guide for Conducting Risk Assessments," NIST Special Publication 800-30, Washington, DC, USA. http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf

[27] Nichols, A. L. & Maner, J. K. (2008). "The good-subject effect: Investigating participant demand characteristics," *Journal of General Psychology* 135(2), 151-166.

[28] Nosek, B. A. & Lakens, D. (2014). "Registered Reports: A Method to Increase the Credibility of Published Results," *Social Psychology* 45, 137-141.

[29] Open Science Collaboration. (2015). "Estimating the reproducibility of psychological science," *Science* 349(6251), 943-952. DOI: 10.1126/science.aac4716

[30] Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). "Instructional manipulation checks: Detecting satisficing to increase statistical power," *Journal of Experimental Social Psychology* 45(4), 867–872. DOI: http://doi.org/10.1016/j.jesp.2009.03.009

[31] Orne, M. T. (1962). "On The Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications," *American Psychologist* 17(11), 776–783.

[32] Paolacci, G., Chandler, J. & Ipeirotis, P. G. (2010). "Running experiments on Amazon Mechanical Turk," *Judgement and Decision Making* 5(5), 411–419.

[33] Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., & Jerram, C. (2015). "The design of phishing studies: Challenges for researchers," *Computers & Security* 52, 194-206. DOI: http://doi.org/10.1016/j.cose.2015.02.008

[34] Pattinson, M., Jerram, C., Parsons, K., McCormac, A. & Butavicius, M. (2012). "Why do some people manage phishing e-mails better than others?" *Information Management & Computer Security* 20(1), 18–28.

[35] Schwartz, D., Fischhoff, B., Krishnamurti, T. & Sowell, F. (2013). "The Hawthorne Effect and energy awareness," *PNAS*, 110(38), 15242-15246.

[36] Sheng, S., Holbrook, M. B., Kumaraguru, P., Cranor, L. F. & Downs, J. (2010). "Who Falls for Phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions," In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI '10*, 1-10.

[37] Sheng, S., Kumaraguru, P., Acquisti, A., Cranor, L. F., & Hong, J. (2009). "Improving phishing countermeasures: An analysis of expert interviews," In *Proceedings of the 4th APWG eCrime Researchers Summit*.

[38] Sotirakopoulos, A., Hawkey, K., & Beznosov, K. (2011). "On the Challenges in Usable Security Lab Studies: Lessons Learned from Replicating a Study on SSL Warnings," In *Proceedings of the Symposium on Usable Privacy and Security: SOUPS '11*.

[39] Symantec Corporation. (2017). *Internet Security Threat Report*. https://www.symantec.com/security-center/threat-report

[40] Verizon. (2017). *2017 Data Breach Investigations Report*. Retrieved from http://www.verizonenterprise.com/verizon-insights-lab/dbir/2017/

[41] Vishwanath, A., Herath, T., Chen, R., Wang, J., Rao, H. R. (2011). "Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model," *Decision Support Systems* 51(3), 576–586. DOI: http://doi.org/10.1016/j.dss.2011.03.002

[42] Wang, J., Herath, T., Chen, R., Vishwanath, A., & Rao, H. R. (2012). "Phishing susceptibility: An investigation into the processing of a targeted spear phishing email," *IEEE Transactions on Professional Communication* 55(4), 345–362. DOI: http://doi.org/10.1109/TPC.2012.2208392

[43] Wash, R. (2010). "Folk Models of Home Computer Security," In *Proceedings of the Symposium on Usable Privacy and Security: SOUPS '10*.

[44] Wash, R. & Rader, E. (2015). "Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users," In *Proceedings of the Symposium on Usable Privacy and Security: SOUPS '15*, 309-325.

[45] Wasserstein, R. L. & Lazar, N. A. (2016). "The ASA's statement on p-values: context, process, and purpose," *The American Statistician* 70(2), 129-133.

[46] Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S. & Kibbi, N. (2007). "Low target prevalence is a stubborn source of errors in visual search tasks," *Journal of Experimental Psychology: General* 136(4), 623–638. DOI: http://doi.org/10.1037/0096-3445.136.4.623

[47] Wright, R. T. & Marett, K. (2010). "The Influence of Experiential and Dispositional Factors in Phishing: An Empirical Investigation of the Deceived," *Journal of Management Information Systems* 27(1), 273–303.

APPENDIX

A. Open Data

The data and code for this paper are available at https://osf.io/6dknx/.