Risk Analysis, Vol. 38, No. 4, 2018 DOI: 10.1111/risa.12917

Setting Priorities in Behavioral Interventions: An Application to Reducing Phishing Risk

Casey Inez Canfield [1,*] and Baruch Fischhoff [1,2]

Phishing risk is a growing area of concern for corporations, governments, and individuals. Given the evidence that users vary widely in their vulnerability to phishing attacks, we demonstrate an approach for assessing the benefits and costs of interventions that target the most vulnerable users. Our approach uses Monte Carlo simulation to (1) identify which users were most vulnerable, in signal detection theory terms; (2) assess the proportion of system-level risk attributable to the most vulnerable users; (3) estimate the monetary benefit and cost of behavioral interventions targeting different vulnerability levels; and (4) evaluate the sensitivity of these results to whether the attacks involve random or spear phishing. Using parameter estimates from previous research, we find that the most vulnerable users were less cautious and less able to distinguish between phishing and legitimate emails (positive response bias and low sensitivity, in signal detection theory terms). They also accounted for a large share of phishing risk for both random and spear phishing attacks. Under these conditions, our analysis estimates much greater net benefit for behavioral interventions that target these vulnerable users. Within the range of the model's assumptions, there was generally net benefit even for the least vulnerable users. However, the differences in the return on investment for interventions with users with different degrees of vulnerability indicate the importance of measuring that performance, and letting it guide interventions. This study suggests that interventions to reduce response bias, rather than to increase sensitivity, have greater net benefit.

KEY WORDS: Behavioral intervention; benefit–cost analysis; phishing; signal detection theory; system-level risk

1. INTRODUCTION

Most cyberattacks begin with a phishing attack via email, or increasingly, social media websites.[1,2] Phishing attacks seek to gather information or trick users into inadvertently installing malware that allows hackers to access networks. Often, attackers mass email employees, gathering information from out-of-office replies and bounce notices, along with whatever information users are tricked into providing. This information can then be used to design attacks, called spear phishing, that use personal information (e.g., known contacts, industry language, and victims' names) to design more realistic and persuasive messages. When successful, phishing attacks may provide hackers with wide access to an organization's network, with their success depending on the organization's internal security practices, the type of account that has been accessed, and the hackers' goal. At present, many firms are trying to reduce phishing vulnerability, as evidenced by the market for anti-phishing training and analytics (e.g., Wombat Security[3] and PhishMe[4]).

[1]Engineering & Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA.
[2]Institute for Politics and Strategy, Carnegie Mellon University, Pittsburgh, PA, USA.
*Address correspondence to Casey Canfield, Engineering & Public Policy, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213-3815, USA; tel: +412-268-2000; caseycan@gmail.com.

When employing such behavioral interventions, organizations want to ensure that they are allocating resources cost effectively. Users' phishing susceptibility can vary widely.[5] Here, we examine the implications of that variation when evaluating the monetary benefits and costs of behavioral interventions that target subgroups varying in their phishing vulnerability. Our proposed approach considers (1) how to identify poor detectors, (2) how to estimate their contribution to overall system vulnerability, and (3) how to assess the benefits and costs of interventions targeting them.

1.1. Modeling Phishing Risk

Cybersecurity risk (R) is often defined as a function of threat (T), vulnerability (V), and impact (I).[6] In this formulation, impact is the cost of a successful attack, in terms of both direct costs (e.g., from a theft) and indirect costs (e.g., from loss of reputation, productivity, safety, etc.). The probability (P) of a successful attack is a function of threat and vulnerability. Threats include malicious attacks (e.g., phishing) and errors (e.g., accidentially publishing private information). Vulnerabilities are human, organizational, and technical weaknesses that can be exploited by an adversary.[7,8] These elements are related symbolically by the following equations:

$$R = I^* P,$$

$$P = F(T, V).$$

It is typically impossible to estimate the absolute value of R. Most notably, the threat is unknown and perhaps varying—in part, as a function of adversaries' perceptions of the vulnerabilities and targets' responses to them. Generally, an organization cannot control threats, but must rely on legal and political authorities for protection. It can, however, try to reduce the impact of attacks (e.g., through network segmentation or limiting permissions across the network) or its vulnerability (e.g., through behavioral interventions). Given the difficulty of estimating absolute risk, interventions are best analyzed in terms of their relative contribution to reducing system risk, holding threat constant. Here, we develop a model for such analyses and illustrate it with measures of users' performance taken from behavioral experiments and measures of intervention effectiveness taken from the research literature. The model considers variation in both user performance and intervention effectiveness. The following section reviews the evidence on both forms of performance (and variation), translating it into analytic terms.

1.2. Accounting for Human Variation

Managing phishing risks is an example of what human factors (or ergonomics) researchers call vigilance tasks, ones in which individuals must monitor their environment for a signal. Mackworth first studied vigilance in 1948 in order to determine the optimal watch length for airborne radar operators, seeking to maximize accuracy in submarine detection.[9]

Since then, vigilance research has identified task, individual, and environmental variables that can affect performance.[10] Task factors include base rate, payoffs, and similarity of stimuli.[11] Studies typically find that people are less likely to identify a signal when there is a low base rate, the cost of missing a signal is low, the cost of mistaking noise for a signal is high, or there is little difference between the signal and noise (e.g., navigating a dimly lit room). Individual factors include experience, personality, and demographics. People may be less likely to identify a signal correctly when they are less experienced, more impulsive, older, or less intelligent.[10] Environmental factors that increase stress, such as uncomfortable ambient conditions or greater workload, can reduce performance.[10] The wide range of such shaping factors leads one to expect variation in performance both within and between users. For example, even highly trained users might occasionally be distracted and fall for phishing attacks, especially when attacks are rare and their workload high. The following sections review results regarding variation in susceptibility to phishing attacks and response to behavioral interventions.

1.2.1. Variation in Phishing Susceptibility

Following vigilance research, we conceptualize human phishing vulnerability in signal detection theory (SDT) terms, with performance measured as sensitivity (d') and response bias (c).[12] Sensitivity refers to users' ability to distinguish between signal and noise, here, phishing and legitimate emails, respectively. Greater sensitivity (as reflected in larger values of d') indicates greater discrimination ability. Response bias refers to users' tendency to treat an email as phishing or legitimate when translating their uncertain beliefs into actions. When response bias (c) is 0, users show no bias. When response bias is

negative, users are biased toward treating emails as phishing; when response bias is positive, users are biased toward treating emails as legitimate.

As with vigilance research, phishing detection research has found that vulnerability (i.e., sensitivity and response bias) can be influenced by task, individual, and environmental factors.[10,13,14] For task factors, Canfield *et al.* found that users were less sensitive and more cautious (i.e., had lower sensitivity and response bias) when asked to choose an action (e.g., click the link), rather than to characterize an email as phishing or not, likely because the perceived consequences (or payoffs) were higher for actions.[15] In addition, users who perceived worse consequences of phishing were more cautious (i.e., had lower response bias). Providing information about the base rate of phishing emails in the test set, however, had no effect on performance. Wolfe and colleagues found sensitivity to base rates, in the context of baggage screening, in studies that manipulated the base rate (rather than just reporting it to users).[16] Spear phishing could be construed as reducing sensitivity by increasing the similarity of phishing and legitimate emails.

For individual factors, Wright and Marett distinguish between experiential and dispositional variables.[14] In terms of experience, users with more computer knowledge tend to be less vulnerable.[5,13,17,18] In terms of disposition, users who are more impulsive, trusting, and risk seeking, tend to be more susceptible.[5,17,19] Most phishing detection research assesses susceptibility in terms of accuracy, rather than in signal detection terms. Extrapolating from vigilance research to phishing, trust is more likely to influence response bias because it would influence users' general inclination to view emails as threatening. Trust would only influence sensitivity if it led users to pay less attention to their emails, hence not fully apply their detection skills.[20] These individual factors can interact with demographic factors; for example, women tend to have less computer knowledge and younger people tend to be less risk averse, both of which may make them more vulnerable.[17]

Environmental factors, such as workload and time pressure, increase stress, which may increase vulnerability. For example, users who receive many emails and open new email as a habit, without much conscious effort, are more vulnerable.[13,21] Similarly, users who multitask while looking at their emails or work under tight time deadlines, encouraging cursory review of emails, might be more vulnerable.

At present, a common way for organizations to evaluate phishing susceptibility is with "embedded training"— sending fake phishing emails to employees, observing who clicks on the links, and (potentially) providing remedial training.[22] An alternative strategy is to use an independent measure of phishing susceptibility to identify users needing extra training or protection. Tests of computer security knowledge or attitudes might guide such targeting.[23] Canfield *et al.* have developed a test of phishing susceptibility that system operators might employ.[15] It characterizes vulnerability in terms of sensitivity and response bias, thereby providing parameter estimates that could be used in quantitative risk analyses. The next section summarizes evidence regarding the effectiveness of interventions that might be administered to some (or all) of a system's users based on their performance.

1.2.2. Effectiveness of Anti-phishing Interventions

Vigilance researchers have long been interested in improving the detection of low base rate phenomena. Typically, these are high-consequence events (e.g., diagnosing cancer, detecting an enemy submarine, and avoiding phishing links), where the cost of missing an event is high, but it is impossible to treat every case as an impending disaster (because signals are so infrequent). For example, one cannot responsibly tell people that they have cancer based on weak signals just to ensure that all cases are caught.[24] Similarly, it is not realistic to treat a large portion of emails as phishing, as that would interfere with users' primary work duties. Given how few emails are phishing, such advice might, at some point, be ignored.[25]

In vigilance research, most interventions focus on task or individual factors. For example, in the context of baggage screening for airport security, Wolfe and colleagues found that exposing operators to brief bursts of training at a high base rate with full feedback reduced response bias, even after returning to a real world with a low base rate and limited feedback.[16] This result suggests that regularly performing such training might encourage observers to maintain a low response bias despite the low base rate.[20] With air-traffic control, Bisseret observed that more experienced controllers had a lower response bias than did new recruits, but did not differ in sensitivity.[26] Such results suggest that experience can reduce the perceived costs of false alarms and encourage reporting.

Table I. Effectiveness of Interventions in the Literature, in Terms of Sensitivity (d') and Response Bias (c)

Reference	Intervention	Task	$\Delta\ d'$	$\Delta\ c$
Kumaraguru *et al.*[22]	Educational materials	Phishing detection	0.62	−0.54
Kumaraguru *et al.*[22]	Embedded training (PhishGuru)	Phishing detection	1.73	−0.52
Kumaraguru *et al.*[22]	Game in lab (Antiphishing Phil)	Phishing detection	1.09	0.00
Kumaraguru *et al.*[22]	Game in field (Antiphishing Phil)	Phishing detection	0.97	0.37
Ben-Asher and Gonzalez[28],[a]	Expertise	Network attacks	0.07	0.06
Wolfe *et al.*[16],[a]	Burst of high base rate with feedback	Baggage screening	−0.49	−0.95
Bisseret[26],[b]	Experience	Air-traffic control	0.02	−0.18
		Average Effect Size	*0.57*	*−0.25*

[a]Reported hit and false alarm rates, converted to d' and c.
[b]Reported as β and converted to c where $c = \ln(\beta)/d'$.
Note: See Stanislaw and Todorov for more details on the calculation of SDT parameters.[29]

For phishing detection, common behavioral interventions include embedded training (feedback on misses), warnings (about known risks), and education (ranging from information to games). Most studies have measured performance in terms of accuracy (i.e., the number of successful attacks in some period of observation). However, accuracy conflates sensitivity and response bias. Accuracy can be increased through better discrimination or more cautious decision rules. In one of the few studies measuring phishing detection performance in SDT terms, Kumaraguru and colleagues found that embedded training increased sensitivity and decreased response bias.[22] Embedded training is similar to the intervention tested by Wolfe and colleagues, but includes feedback only on false negatives (i.e., cases where phishing attacks are missed).[16]

Interventions that increase attention or effort have sometimes been found to increase sensitivity. For example, Parsons and colleagues found that telling users that they were being evaluated for their phishing detection ability increased their sensitivity without changing their response bias.[27] Wolfe and colleagues observed an increase in sensitivity during high base rate training trials.[20] However, unlike the sustained change observed with response bias, sensitivity returned to the previous value immediately after training ended. One possible explanation is that screeners could not sustain the heightened level of attention that they mustered during the training. This result suggests that it may be better to focus on interventions that influence response bias, rather than sensitivity.

Table I and Fig. 1 summarize studies of behavioral interventions that reported results in SDT terms. They were identified by using the joint search terms of "signal detection theory" and "behavioral intervention" in Google Scholar, which produced 76 papers. We identified an additional 65 papers by using the joint search terms "signal detection theory," "phishing," and "experiment." We then eliminated papers that did not report empirical evidence of evaluating a behavioral intervention in SDT terms. That left seven studies in four articles. We use these few studies for their SDT parameter estimates, recognizing that they constitute a fraction of the studies evaluating behavioral interventions. The analysis demonstrated here suggests the value of estimating the effects of interventions in SDT terms.

Fig. 1 contrasts sensitivity and response bias for these studies, before and after the intervention. In this small sample of studies, the interventions were more effective at improving sensitivity for phishing detection (black circles), compared to the other contexts (blue squares), while having similar effects on response bias. For improving sensitivity, the most effective intervention was embedded training. For decreasing response bias, a burst of high base rate training with feedback was most effective. Few studies reported individual variation in intervention effectiveness. However, given the heterogeneity of baseline performance, it seems plausible that interventions might not influence all users equally. Our analysis allows for this possibility.

Although the vulnerability of a system is determined by its users' sensitivity and response bias, system operators may be concerned about the absolute number of successful attacks. That rate will partially determine the total cost to their system from attacks and the appropriate investment in their reduction.

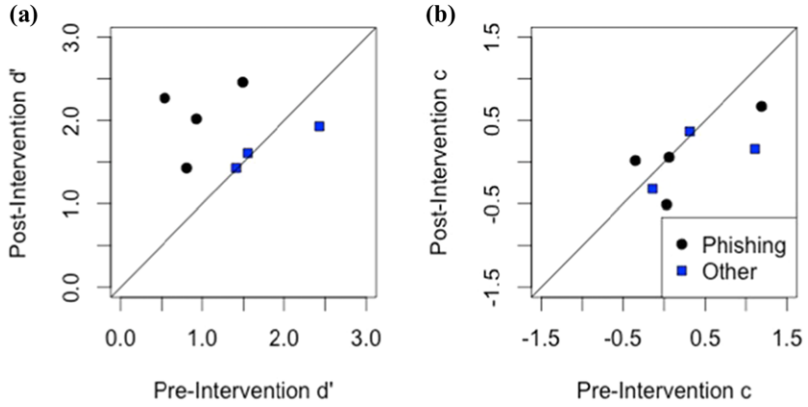

Fig. 1. Average change in (a) sensitivity or d' and (b) response bias or c for various behavioral interventions.

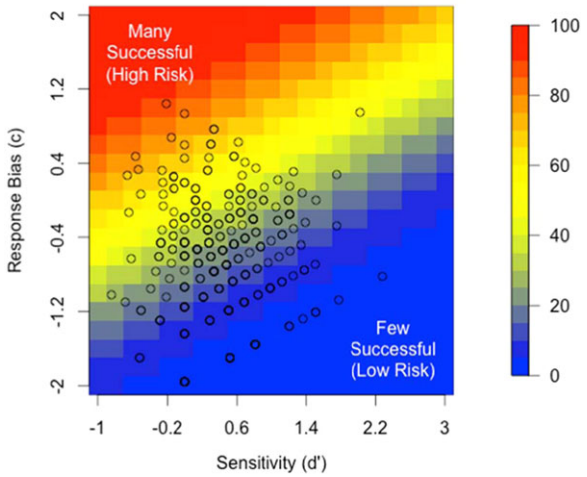

Fig. 2. Number of successful phishing attacks out of 100 (denoted by color) as a function of sensitivity (d') and response bias (c). Observations from Canfield *et al.* are plotted in black.[15] Risk is high when sensitivity is low and users are biased toward clicking on links in emails (positive response bias).

Here, we assess performance in terms of the number of successful phishing attacks (of each 100 attempts). Fig. 2 shows performance for different values of sensitivity and response bias. When response bias is negative and sensitivity is high, the risk is low (blue, bottom-right corner). When response bias is positive and sensitivity is low, the risk is high (red, top-left corner). As seen in the figure, users can have the same number of successful attacks with varying SDT parameters. For example, a user with $d' = 1.25$ and $c = 0.31$ has the same number of successful attacks as a user with $d' = 0$ and $c = -0.32$.

The black circles in Fig. 2 show the vulnerability of individual participants in Canfield *et al.* as determined by their sensitivity and response bias values.[3][15] The risk model in the next section assesses the value of behavioral interventions for users in a population with this distribution of vulnerability levels (analogous to the color bands in Fig. 2). The model defines benefits in terms of reduced vulnerability (i.e., a lower rate of successful attacks) and costs in terms of those associated with any increase in the rate of false alarms, thereby reducing users' ability to do their jobs (and possibly reducing the intervention's effectiveness over time, as those costs mount), and those associated with implementing behavioral interventions.

The model accommodates the natural variation in phishing susceptibility and the fact that users can have the same level of vulnerability for different reasons (i.e., combinations of sensitivity and response bias), as seen in Fig. 2. It also accommodates the fact that interventions can have different effects on the two SDT parameters, as seen in Table I. As a result, interventions can have different effects on users with the same vulnerability. We use a simulation to (1) identify poor detectors, defined as the bottom 10% of users; (2) determine the cumulative contribution of those poor detectors to overall system vulnerability; and (3) compare the benefit–cost of behavioral interventions when focused on poor detectors or all users. For the purposes of the present demonstration, we defined "poor detectors" as the bottom 10% of users. In its 2016 Data Breach Investigations Report, Verizon reported that 13% of people tested clicked

[3]Canfield *et al.* estimated sensitivity and response bias for a detection task, "Is this a phishing email?" (Yes/No), as well as a behavior task, "What would you do if you received this email?" (multiple choice).[15] We use the estimates of sensitivity and response bias from the behavior task, which captures the actions affecting system performance better than the detection task. The data are publicly available at https://osf.io/7bx3n/.

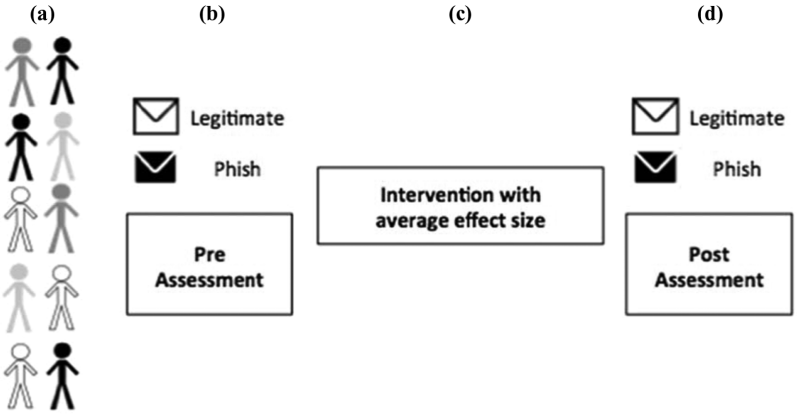

Fig. 3. Overview of risk model.

on a phishing attachment, suggesting a most vulnerable population of about that size.[2] An interested organization could apply the present approach with estimates of these performance parameters for members of its staff.

2. METHOD

The code for this analysis is publicly available at https://osf.io/hkp9a/.

2.1. Overview of Risk Simulation

The present model simulates the effects of behavioral interventions on users' phishing susceptibility for two different types of attacks: random (with no special recognition of the target) and spear phishing (with some personal information). As depicted in Fig. 3, in each iteration of the model, we first generate a sample of individuals with varying vulnerability, defined by sensitivity and response bias, with values drawn from the distribution of empirical estimates reported by Canfield *et al.* (Step A).[15] We then estimate each user's initial (or baseline) performance, in terms of the number of phishing emails that he or she falls for (misses) and the number of legitimate emails that the user mistakes for phishing (false alarms) (Step B). For each user, we sample an intervention from a normal distribution defined by the mean and standard deviation of results from the literature review (in Table I), pooling phishing and non-phishing interventions (Step C). That distribution is used to reflect the variation in the effects of these interventions on individual users (which is not routinely reported in studies). We then recalculate each user's sensitivity and response bias, incorporating the intervention's effects (Step D).

In Steps B and D, we estimate vulnerability separately for random and spear phishing attacks. The ability to detect random phishing attacks is determined by users' initial sensitivity and response bias, plus the effects of any intervention. Because spear phishing emails are designed to look like legitimate emails, users have a lower sensitivity (d'). (Kaivanto adopts a similar approach.[30]) The extent of that reduction in sensitivity depends on how well the spear phishing email is crafted. As a placeholder for empirical estimates, the model uses a difficulty factor, f, ranging from 0, for a spear phishing attack that is impossible to detect, to 1, for one that is no more difficult than a random phishing attack to detect. In the simulations reported here, the value for f is sampled from a uniform distribution over [0,1].

We assess performance on each simulated email as a draw from a Bernoulli distribution with P_M as the probability of falling for a phishing email (false negative) and P_{FA} as the probability of mistaking a legitimate email for phishing (false positive). This procedure is repeated for each email, both phishing and legitimate, that a user receives. In the simplest scenario (i.e., no interventions or spear phishing attacks), P_M is a function of initial vulnerability (sensitivity and response bias):

$$P_M = 1 - \Phi(0.5d' - c),$$

where Φ represents a standard normal distribution that converts a z-score to a probability.[12] In a scenario with an intervention having estimated impacts $\Delta_{d'}$ and Δ_c, and a spear phishing difficulty factor f, P_M is:

$$P_M = 1 - \Phi\{0.5[(d' + \Delta_{d'})f] - (c + \Delta_c)\}.$$

These variables are summarized in Table II, along with the source of the parameter values used

Table II. Model Inputs

Inputs	Value	Description
Difficulty factor	$f \sim$ Uniform(0,1)	For random phishing attacks, $f = 1$. For spear phishing attacks, f ranges from 0, which eliminates d', to 1, which preserves it.
Sensitivity	$d' \sim$ Normal(0.4, 0.5)	Estimated from experimental data available at https://osf.io/7bx3n/.[13]
Response bias	$c \sim$ Normal(−0.6, 0.65)	Estimated from experimental data available at https://osf.io/7bx3n/.[13]
Effect on d'	$\Delta_{d'} \sim$ Normal(0.57, 0.76)	The mean and standard deviation are based on the literature review (see Table I).
Effect on c	$\Delta_c \sim$ Normal(−0.25, 0.45)	The mean and standard deviation are based on the literature review (see Table I).

in the simulation. We report users' vulnerability (i.e., their probability of falling for a phishing attack, P_M) by decile to facilitate comparing low- and high-performing users. Users in a low decile have a high probability of falling for attacks, while users in a high decile have a low probability (in effect, going down and to the right in Fig. 2). We assume a 1% base rate, so for every phishing email, there are 99 legitimate emails. We estimate false alarms per user as P_{FA}:

$$P_{FA} = \Phi\{-0.5[d' + \Delta_{d'}] - (c + \Delta_c)\}.$$

We report performance (or phishing accuracy) in terms of the rate of phishing emails that are missed. High-performing users fall for many attacks, while low-performing users fall for few. We make a distinction between expected vulnerability (estimated probability based on sensitivity and response bias) and observed performance (simulated as draws from a Bernoulli distribution) in order to recognize that observations may not perfectly reflect reality (as defined by sensitivity and response bias).

We use a Monte Carlo simulation to incorporate uncertainty by assigning a distribution to each parameter. The results represent the outcome of 1,000 iterations, each involving 100 phishing attacks against 100 users with a 1% base rate. We compare different scenarios in terms of their benefit–cost, as described in the following section.

2.2. Benefit–Cost Analysis

Benefit–cost analysis is a systematic, analytical approach for assessing tradeoffs among options.[31] Here, we estimate the difference between the benefits and costs (net benefit) of anti-phishing interventions. We include both the direct cost of the intervention (e.g., usage fees, lost time, and productivity) and indirect costs from changed behavior (e.g., increased false alarms). We measure the impact of the intervention as the change in the number of successful attacks and false alarms. When that change is negative, an intervention reduces successful attacks and false alarms enough to provide net benefits.

The cost of successful attacks could be as low as that of having to change a compromised password or as high as a major data breach. The cost of a false alarm could be as low as typing a URL into a browser (rather than clicking on the link) or as high as a lost business opportunity. We assume that probabilities are not uniform across the range of possible impacts, but that high-cost events are rare. Therefore, we model the costs of attacks and false alarms with a lognormal distribution, which has a long positive tail, to accommodate those rare, high-cost events. Table III summarizes these assumptions.

3. RESULTS AND DISCUSSION

The results are presented in three sections. Section 3.1 compares observed performance and expected vulnerability. Section 3.2 assesses the cumulative vulnerability by decile of individual users' vulnerability. Section 3.3 examines the costs and benefits of behavioral interventions for different deciles of users.

3.1. Measurement of Vulnerability

This section compares observed performance to expected phishing vulnerability, (P_M), in order to identify the characteristics of poor detectors. We define a distribution of users, characterized by their relative proficiency as detectors (in deciles), as potential targets of selective interventions. We use the values of sensitivity and response bias observed by Canfield *et al.*[15]

Fig. 4(a) shows the performance of users (in terms of phishing accuracy, defined as the percent of

Table III. Summary of Assumptions for Benefit–Cost Analysis

	Cost	Benefit	Value	Source
Attack	Additional successful attacks	Avoided attacks	1,800 × Lognormal(0,1)	Cyveillance[32]
False alarm	Additional false alarms	Avoided false alarms	Lognormal(0,2)	
Intervention	Cost of implementation; lost productivity	N/A	Uniform(1,10); Uniform(10,100)	Ponemon Institute;[33] range accounts for time spent on intervention (1–60 minutes), frequency (1–52 times/year), and hourly wage for professionals ($20–50).

phishing emails avoided) at each vulnerability decile. The means are monotonically related, by definition. Their slope proves to be relatively linear for the highest deciles (with the fewest successful attacks). However, they spread out for the lowest deciles, suggesting the potential value of targeting the poorest detectors. The measure of phishing accuracy in Fig. 4(a) assumes a test with 100 phishing emails. Figs. 4(b) and 4(c) show the deciles of vulnerability as a function of the two SDT parameters separately. As would be expected, users in the 10th decile have relatively high sensitivity and negative response bias, while users in the first decile have low sensitivity and positive response bias. For each decile, the range is wider for sensitivity than for response bias, as reflected in a stronger correlation between P_M and response bias, $r(98) = 0.91$, $p < 0.001$, than with sensitivity, $r(98) = -0.38$, $p < 0.001$. This suggests that response bias is a more influential parameter than sensitivity for interventions (across all users). This result follows from the effect of response bias being twice that of sensitivity in the equation for vulnerability (P_M).

3.2. Cumulative Vulnerability by Decile

Second, we assess vulnerability by decile, as a basis for evaluating the potential benefit of targeting poor detectors for behavioral interventions. Fig. 5(a) translates the estimates of Fig. 4(a) into the percentage of successful attacks per decile. The black circles show these estimates for random phishing attacks,

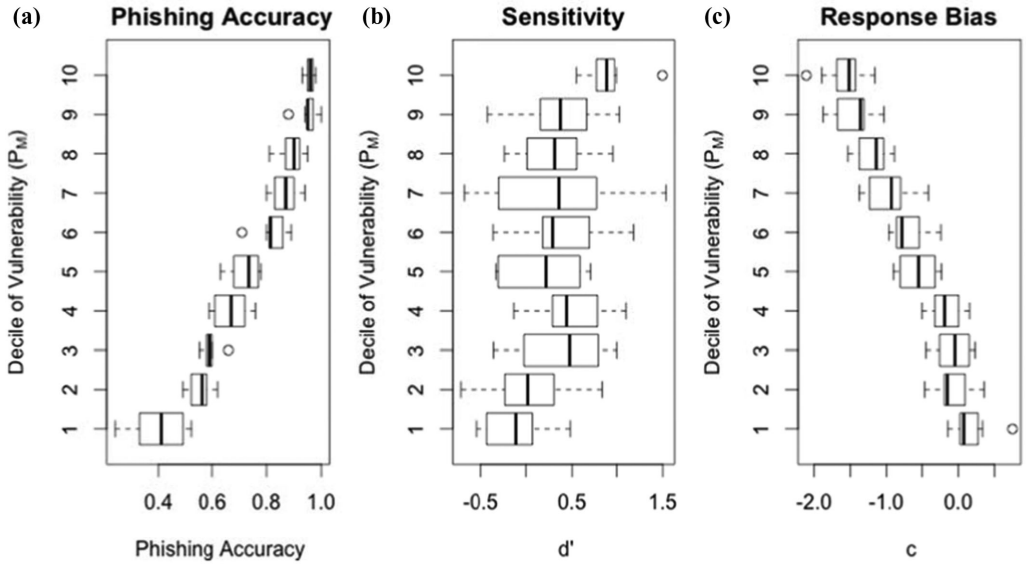

Fig. 4. Decile of probability of falling for an attack as a function of (a) performance (accuracy) for 100 phishing emails, (b) sensitivity [d'], and (c) response bias [c].

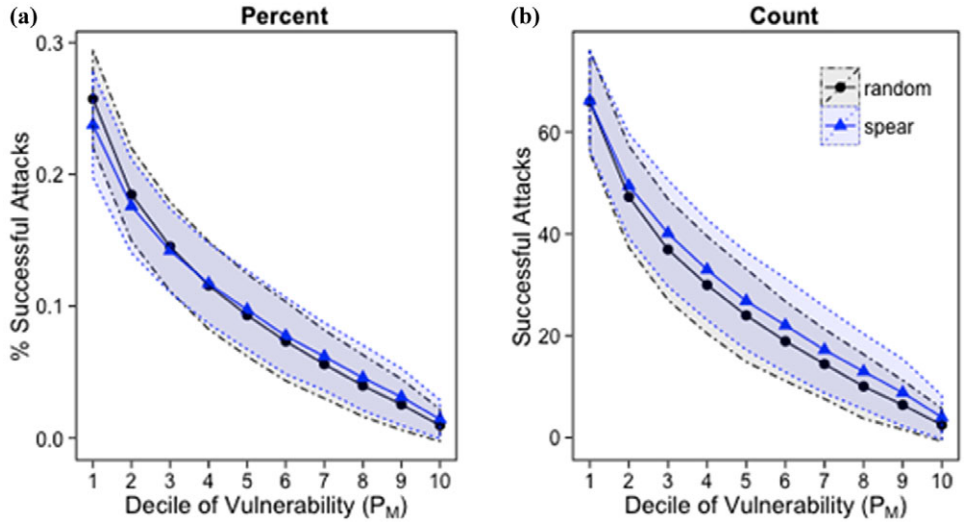

Fig. 5. (a) Percent and (b) count of successful attacks for 1 attack on 1,000 users, by vulnerability decile. The error bars are ± 2 standard deviations.

where the bottom 10% of users account for 26% of the total number of successful attacks. The blue triangles are for spear phishing attacks, which reduce sensitivity, with the bottom 10% of users accounting for 24% of successful attacks. The two curves are similar, despite the greater difficulty of distinguishing spear phishing attacks, because of the relatively weak relationship between sensitivity and vulnerability (Fig. 4b). Fig. 5(b) shows the necessarily similar pattern for the count of successful attacks. In summary, poor detectors (bottom 10%) account for a disproportionate share of the modeled organization's overall vulnerability for both random and spear phishing. This suggests that it may be worthwhile to focus intervention resources on poor detectors. That question is addressed in the next section.

3.3. Benefit–Cost Analysis of Behavioral Interventions

Third, we evaluate the benefit–cost of behavioral interventions. The benefits of an intervention are determined by the net reduction in the numbers of successful attacks and legitimate emails mistaken as phishing (false alarms). The costs of an intervention include those associated with its implementation (e.g., fees, lost productivity) and any additional successful attacks and false alarms that it unintentionally creates (e.g., by increasing trust in spam filters, which do not completely protect users). We first report results from a Monte Carlo simulation varying the type of attack, looking at the net benefit of an intervention administered to users in each decile. We then report a sensitivity analysis examining the influence of our assumptions.

Fig. 6 shows the net benefit of interventions, when applied to users in each decile, for random and spear phishing attacks. Given the fixed costs of the intervention (per user), the net benefits are much greater for users in the lower deciles, who contribute a disproportionate share of the system's vulnerability (Fig. 5). However, some benefit exists even with the best detectors. For low-decile users, the net benefit is somewhat greater for random attacks because they are easier to detect, so interventions have a larger effect. Because low-decile users fall for more attacks overall, the difference is larger. For random attacks, the mean net benefit is $580,000 (SD = $220,000) for users in the first performance decile, equal to 20% of the total net benefit of a systemwide program. It is $56,000 (SD = $50,000) for users in the 10th decile, or just 2% of the total net benefit. For spear phishing attacks, the mean net benefit is $440,000 (SD = $180,000) for users in the first performance decile, or 18% of the total benefit. For users in the 10th decile, the mean net benefit is $60,000 (SD = $48,000) or 2% of the total net benefit. Thus, the net benefit is positive (above the dotted line in Fig. 6), under most conditions for all users. The next section performs sensitivity analyses, varying model parameters.

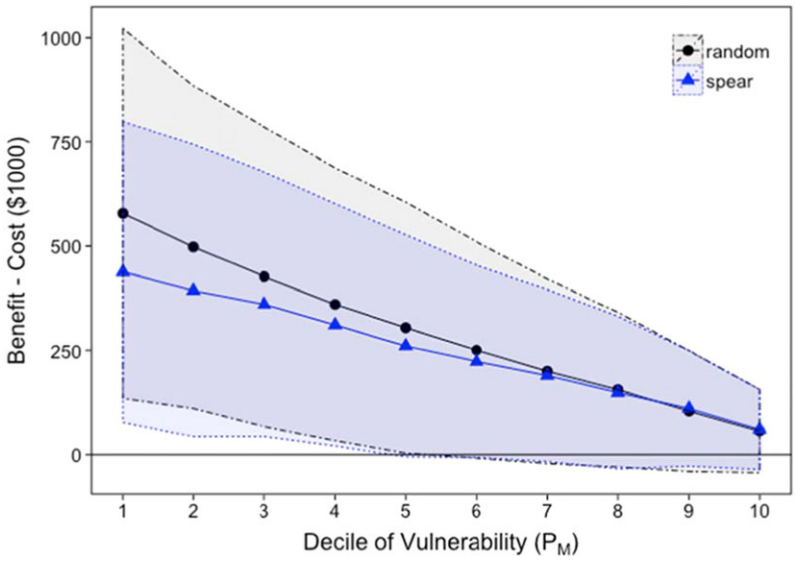

Fig. 6. Benefit–cost by vulnerability decile, where scenarios above the 0 line have positive net benefit.

Table IV. Percent Change of Mean Benefit–Cost for Random Attacks from the Baseline Scenario (Reported for the 1st and 10th Deciles)

			1st	10th		1st	10th
Parameter	Baseline	Worst	B–C	B–C	Best	B–C	B–C
1. Mean d'	0.4	0	−4%	43%	3	−46%	−85%
2. Mean c	−0.6	2	−97%	820%	−2	−51%	−98%
3. Effect on d'	0.57	−1	**−170%**	**−290%**	1	46%	64%
4. Effect on c	−0.25	1	**−170%**	−73%	−1	92%	−75%
5. Cost of Attack	$3,000	$0	−99%	−94%	$200,000	6,500%	6,100%
6. Cost of FA	$7	$0	−1%	−9%	$100,000	18,000%	100,000%
7. Cost of Intervention	$60	$10,000	−16%	**−160%**	$10	−1%	0%
1st B–C	$610,000						
10th B–C	$63,000						

Note: Each parameter is varied independently, while the other parameters are held at the baseline value. Cases where the change in net benefit is less than −100% (bolded) are where the benefit–cost crosses 0.

3.3.1. Sensitivity Analysis

Table IV shows the sensitivity of these estimates to varying each model parameter independently, for users in each decile, expressed as percent change from baseline performance (without the intervention). Each row represents a parameter that was varied. The baseline assumptions are the mean inputs from the Monte Carlo model. The worst and best scenario assumptions are either the minimum or maximum inputs from the Monte Carlo model or other values of interest, as noted in the text below. The first column in Table IV provides the baseline assumptions, yielding a net benefit of $610,000 for the first decile and $63,000 for the 10th decile. Because the results are reported in terms of percent change of net benefit from the baseline, it is less than −100% (bolded in Table IV) where benefit–cost crosses 0.

Row 1 shows the effects of varying the mean sensitivity of users across the nominal range of sensitivity (0–3). The baseline scenario used 0.4, the mean sensitivity observed in the behavior task by Canfield *et al.*[15] When the initial mean sensitivity of users is very poor ($d' = 0$), high-decile users benefit from an intervention more than low-decile users because they can be more responsive (because they already have some, rather than no, ability to detect phishing emails). Even with the strongest of interventions, low-decile users' sensitivity still reflects weak discrimination. When the initial mean sensitivity of users is very high ($d' = 3$), interventions have limited

net benefit for high-decile users, who are already able to distinguish almost perfectly between phishing and legitimate emails.

Row 2 varies the mean response bias (c) of users across the nominal range of response bias (-2 to 2). The baseline scenario used -0.6, the mean response bias observed in the behavior task by Canfield *et al.*[15] For the worst-case value, where users are not very cautious ($c = 2$), the benefits are much greater for high-decile users. Low-decile users are so incautious to begin with (at the baseline value) that the intervention still leaves them falling for many attacks. At the best-case value, where users are already very cautious ($c = -2$), the intervention has little net benefit for all users.

Rows 3 and 4 show the results of sensitivity analyses varying the effectiveness of interventions. The worst-case values were chosen to represent interventions that not only failed, but backfired, significantly reducing sensitivity or increasing response bias ($\Delta_{d'} = -1$, $\Delta_c = 1$). They have net costs (rather than benefits). The best-case values were chosen to represent very effective interventions, decreasing or increasing sensitivity and response bias by 1. They lead to increased net benefits. Interventions that increase sensitivity ($\Delta_{d'} = 1$) provide greater net benefit for all users. Interventions that decrease response bias ($\Delta_c = -1$) increase net benefit for low-decile users, but decrease net benefit for high-decile users (due to increased false alarms). Thus, interventions can backfire if they reduce detection performance (by decreasing sensitivity or increasing response bias) or increase false alarms (by decreasing response bias too much for high-decile users).

The final three rows vary the financial costs of successful attacks, false alarms, and interventions. We assessed the worst- and best-case values used in the Monte Carlo model. The one exception is for the worst intervention cost: a $10,000 intervention was not used in the Monte Carlo model, but represents the minimum cost for the net benefit to be negative. This is, of course, a very unrealistic cost and emphasizes the cost effectiveness of interventions. Very expensive attacks (e.g., costing $200,000/user affected) and false alarms (e.g., costing $100,000 per email) make any behavioral interventions extremely cost effective because interventions generally reduce successful attacks and false alarms. However, such extreme events may be so unusual that they are not worth considering. If there is no cost of an attack, then there is little net benefit, except from avoided false alarms. If there is no cost of a false alarm, then the net benefit does not change, meaning that most of the estimated net benefit can be attributed to avoided attacks. The cost of the intervention outweighs the benefits for high-decile users when it approaches $10,000 per person. These results are summarized in Table IV.

4. CONCLUSION

We demonstrate an approach, using a Monte Carlo simulation, to assess the value of implementing anti-phishing behavioral interventions under a wide range of scenarios. Our approach has three steps. First, identify poor detectors, defined here as the bottom 10% (or first decile). Second, assess system vulnerability due to poor detectors. Last, perform benefit–cost analyses, considering the sensitivity of estimates to modeling assumptions. The results of these analyses indicate the value of (re)allocating resources to focus on poor detectors, rather than trying to reduce the susceptibility of all users. Doing that requires identifying those poor users. Canfield *et al.* designed a test to do that, quantifying individual performance in SDT terms suited to system analysis.[15] Although there is evidence of construct validity (specifically for the behavior task response bias), attempts to use behavioral data to assess predictive validity of this test were inconclusive.[34] In the present demonstration of the approach, we use estimates of individual performance from that study, estimates of the effectiveness of behavioral interventions from (the relatively few) studies formulated in SDT terms, and estimates of benefits and costs from the professional literature.

For the modeled situation, our analyses had three primary findings.

First, poor detectors tend to have both low sensitivity and high response bias (indicating that they treat most emails as legitimate). Of the two parameters, response bias is much more closely related to vulnerability, suggesting that interventions should focus on response bias. Under normal operating conditions, without a large set of observations from embedded training, it is difficult to tell whether users who perform poorly are bad at detecting phishing emails or are simply unlucky enough to have been a victim in that sample of observations. The test by Canfield *et al.* was designed to provide performance estimates for situations where system operators cannot collect appropriate data under normal operating conditions.[15]

Second, poor detectors create a disproportionate share of the overall risk—by definition. The simulation estimates just how great that share is. Under

the model assumptions, it is quite large, suggesting the potential value of targeting them for interventions. The analyses find it to be similar for random and spear phishing attacks, reflecting the finding that sensitivity (which reflect the type of attack) is only weakly related to vulnerability.

Third, the benefit–cost analysis shows the value of focusing resources on the more susceptible users—although there is some net benefit with almost all users. Interventions may have a negative net benefit if they increase false alarms (beyond the benefit of avoided attacks) or inadvertently decrease sensitivity or increase response bias. For example, a spam filter might increase response bias if users believe that they no longer need to watch out for phishing emails because the filter will catch them.

Once an organization has identified whom to target, the present analyses indicate that it would have greater net benefits from interventions to reduce response bias than from interventions aimed at increasing sensitivity. The relatively few studies measuring performance in these terms suggests that interventions focused on response bias have both greater immediate impact and longer-lasting effects. Embedded training appears to reduce response bias by leading users to see phishing as more likely and more consequential.[22] In addition, because it provides feedback only on misses (false negatives), users only receive the intervention when their response bias has shifted to be more positive and they are tricked by the training email. In contrast, interventions like AntiPhishing Phil, a game that teaches users how to identify suspicious URLs, appears primarily to affect sensitivity, hence may be less effective, even with similar changes in accuracy.[22]

One limit to the model is that it omits some potentially relevant costs of behavioral interventions, such as employee opposition, thereby increasing its estimates of net benefit. In a study of phishing using social connections, Jagatic and colleagues faced strong criticism for using real names as senders of fake phishing emails.[35] An organization could face similar issues when attempting to train users to detect attacks. Caputo and colleagues report that some of their users felt ashamed about clicking on embedded training.[36] Users may just be annoyed if the intervention is time consuming or boring.[37] There may also be benefits not in the model, such as increased reporting of phishing emails.

A second limit to the model's application is the parameter estimates. Those for baseline behavior were drawn from an experimental study, rather than direct observation.[15] Those for the effectiveness of interventions were drawn from a literature review that yielded only a few studies with usable values. Those for costs reflected assumptions about organizational conditions. Wider use of an SDT framework in future research could help organizations find parameter estimates better fitting their circumstances.

A third limit, and topic for future research, is the simple way in which we modeled spear phishing—as a uniform distribution over possible reductions in sensitivity. Vigilance research treats difficulty as a function of similarity.[11] That is consistent with the common attacker strategy of making phishing messages as similar to real ones as possible. Understanding how recipients make those similarity judgments would improve these estimates—and perhaps suggest training options. Kaivanto modeled spear phishing in terms of "match quality," a binary factor indicating a fixed reduction in sensitivity, rather than a variable one, as modeled here.[30] In principle, spear phishing messages might also influence response bias, for example, by creating a sense of urgency or tapping into human emotions, such as greed, potentially leading users to lower their threshold for treating email as legitimate.[13]

Although one might also attempt to predict performance based on personal characteristics, such as education, gender, or cognitive style, studies adopting this strategy have had limited success.[15] It seems more promising to focus on understanding the situational determinants of performance, then use those results to extrapolate from situations where behavior has been studied to other situations of interest. In that light, our results suggest that poor detectors may benefit most from interventions designed to reduce their response bias, whereas better detectors may benefit most from interventions focused on spear phishing, which undermines their otherwise greater sensitivity. There is growing interest in the security community in tailoring behavioral interventions.[23] A model like ours can help to direct resources, once performance differences are identified. That model reflects the broader strategy motivating the present research, translating behavioral research into terms that allow risk analyses to evaluate the needs and opportunities for improving system performance.

ACKNOWLEDGMENTS

Casey Canfield was supported by a National Science Foundation Graduate Research Fellowship

(1121895) and the Carnegie Mellon Bertucci Fellowship. We thank Wändi Bruine de Bruin, Stephen Broomell, Lorrie Cranor, Alex Davis, and the Carnegie Mellon Behavior, Decision, and Policy Working Group for their advice and feedback.

REFERENCES

1. Symantec Corporation. Internet Security Threat Report. Mountain View, CA, 2016. Available at: https://www.symantec.com/security-center/threat-report, Accessed July 22, 2016.
2. Verizon. 2016 Data Breach Investigations Report. New York, NY: Verizon, 2016. Available at: http://www.verizonenterprise.com/verizon-insights-lab/dbir/2016/, Accessed July 22, 2016.
3. Wombat Security [Homepage on the Internet]. Pittsburgh: Wombat Security, c2017. Available at: https://www.wombatsecurity.com, Accessed April 16, 2017.
4. PhishMe [Homepage on the Internet]. Leesburg: PhishMe, c2017. Available at: https://phishme.com, Accessed April 16, 2017.
5. Pattinson M, Jerram C, Parsons K, McCormac A, Butavicius M. Why do some people manage phishing e-mails better than others? Information Management & Computer Security, 2012; 20(1):18–28.
6. National Institute for Standards and Technology. Guide for Conducting Risk Assessments. Washington, DC: NIST, 2012. NIST Special Publication 800-30. Available at: http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf, Accessed July 22, 2016.
7. Sun L, Srivastava RP, Mock TJ. An information systems security risk assessment model under the Dempster-Shafer theory of belief functions. Journal of Management Information Systems, 2006; 22(4):109–142.
8. Werlinger R, Hawkey K, Beznosov K. An integrated view of human, organizational, and technological challenges of IT security management. Information Management & Computer Security, 2009; 17(1):4–19.
9. Mackworth NH. The breakdown of vigilance during prolonged visual search. Quarterly Journal of Experimental Psychology, 1948; 1(1):6–21.
10. Ballard JC. Computerized assessment of sustained attention: A review of factors affecting vigilance performance. Journal of Clinical and Experimental Neuropsychology, 1996; 18(6):843–863.
11. Lynn SK, Barrett LF. "Utilizing" signal detection theory. Psychological Science, 2014; 25(9):1663–1673.
12. Macmillan NA, Creelman DC. Detection Theory: A User's Guide. Cambridge: Cambridge University Press, 2004.
13. Vishwanath A, Herath T, Chen R, Wang J, Rao HR. Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. Decision Support Systems, 2011; 51(3):576–586.
14. Wright RT, Marett K. The influence of experiential and dispositional factors in phishing: An empirical investigation of the deceived. Journal of Management Information Systems, 2010; 27(1):273–303.
15. Canfield CI, Fischhoff B, Davis A. Quantifying phishing susceptibility for detection and behavior decisions. Human Factors, 2016; 58(8):1158–1172.
16. Wolfe JM, Horowitz TS, Van Wert MJ, Kenner NM, Place SS, Kibbi N. Low target prevalence is a stubborn source of errors in visual search tasks. Journal of Experimental Psychology: General, 2007; 136(4):623–638.
17. Sheng S, Holbrook MB, Kumaraguru P, Cranor LF, Downs J. Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. Pp. 1–10 in Proceedings of CHI. 2010.
18. Wang J, Herath T, Chen R, Vishwanath A, Rao HR. Phishing susceptibility: An investigation into the processing of a targeted spear phishing email. IEEE Transactions on Professional Communication, 2012; 55(4):345–362.
19. Welk AK, Hong KW, Zielinska OA, Tembe R, Murphy-Hill E, Mayhorn CB. Will the "phisher-men" reel you in? International Journal of Cyber Behavior, Psychology and Learning, 2015; 5(4):1–17.
20. Wolfe JM, Brunelli DN, Rubinstein J, Horowitz TS. Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. Journal of Vision, 2013; 13(3):1–9.
21. Vishwanath A. Examining the distinct antecedents of e-mail habits and its influence on the outcomes of a phishing attack. Journal of Computer-Mediated Communication, 2015; 20(5):570–584.
22. Kumaraguru P, Sheng S, Acquisti A, Cranor LF, Hong J. Teaching Johnny not to fall for phish. ACM Transactions on Internet Technology, 2010; 10(2):1–31.
23. Egelman S, Peer E. The myth of the average user. Pp. 1–13 in Proceedings of the New Security Paradigms Workshop. 2015.
24. Welch HG, Schwartz LM, Woloshin S. Overdiagnosed: Making People Sick in the Pursuit of Health. Boston: Beacon, 2011.
25. Wickens CD, Rice S, Keller D, Hutchins S, Hughes J, Clayton K. False alarms in air traffic control conflict alerting: Is there a "cry wolf" effect? Human Factors, 2009; 51:446–462.
26. Bisseret A. Application of signal detection theory to decision making in supervisory control: The effect of the operator's experience. Ergonomics, 1981; 24(2):81–94.
27. Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., Jerram, C. The design of phishing studies: Challenges for researchers. Computers & Security, 2015; 52:194–206.
28. Ben-Asher N, Gonzalez C. Effects of cyber security knowledge on attack detection. Computers in Human Behavior, 2015; 48(C):51–61.
29. Stanislaw H, Todorov N. Calculation of signal detection theory measures. Behavior Research Methods, Instruments, & Computers, 1999; 31(1):137–149.
30. Kaivanto K. The effect of decentralized behavioral decision making on system-level risk. Risk Analysis, 2014; 34(12):2121–2142.
31. Lave LB. Benefit-Cost Analysis. Pp 104–134 in Hahn RW, (ed). Do the Benefits Exceed the Costs? New York: Oxford University Press, 1996.
32. Cyveillance. The Cost of Phishing: Understanding the True Cost Dynamics Behind Phishing Attacks. 2015. Available at: http://info.cyveillance.com/rs/cyveillanceinc/images/CYV-WP-CostofPhishing.pdf, Accessed July 22, 2016.
33. Ponemon Institute. The Cost of Phishing & Value of Employee Training. 2015. Available at: https://info.wombatsecurity.com/hubfs/Ponemon_Institute_Cost_of_Phishing.pdf, Accessed July 22, 2016.
34. Canfield CI, Davis A, Fischhoff B, Forget A, Pearman S, Thomas J. Replication: Challenges in using data logs to validate phishing detection ability metrics. Pp. 271–284 in Proceedings of the Thirteenth Symposium on Usable Privacy and Security. 2017.
35. Jagatic TN, Johnson NA, Jakobsson M, Menczer F. Social phishing. Communications of the ACM, 2007; 50:94–100.
36. Caputo DD, Pfleeger SL, Freeman JD, Johnson ME. Going spear phishing: Exploring embedded training and awareness. IEEE Security & Privacy, 2014; 12:28–38.
37. Herley C. More is not the answer. IEEE Security & Privacy, 2014; 12:14–19.