Quantifying Phishing Susceptibility for Detection and Behavior Decisions

Casey Inez Canfield, **Baruch Fischhoff**, and **Alex Davis**,
Carnegie Mellon University, Pittsburgh, Pennsylvania

Objective: We use signal detection theory to measure vulnerability to phishing attacks, including variation in performance across task conditions.

Background: Phishing attacks are difficult to prevent with technology alone, as long as technology is operated by people. Those responsible for managing security risks must understand user decision making in order to create and evaluate potential solutions.

Method: Using a scenario-based online task, we performed two experiments comparing performance on two tasks: *detection*, deciding whether an e-mail is phishing, and *behavior*, deciding what to do with an e-mail. In Experiment 1, we manipulated the order of the tasks and notification of the phishing base rate. In Experiment 2, we varied which task participants performed.

Results: In both experiments, despite exhibiting cautious behavior, participants' limited detection ability left them vulnerable to phishing attacks. Greater sensitivity was positively correlated with confidence. Greater willingness to treat e-mails as legitimate was negatively correlated with perceived consequences from their actions and positively correlated with confidence. These patterns were robust across experimental conditions.

Conclusion: Phishing-related decisions are sensitive to individuals' detection ability, response bias, confidence, and perception of consequences. Performance differs when people evaluate messages or respond to them but not when their task varies in other ways.

Application: Based on these results, potential interventions include providing users with feedback on their abilities and information about the consequences of phishing, perhaps targeting those with the worst performance. Signal detection methods offer system operators quantitative assessments of the impacts of interventions and their residual vulnerability.

Keywords: signal detection theory, cybersecurity, vigilance, perception-action, metacognition

Address correspondence to Casey Inez Canfield, Engineering and Public Policy, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA; e-mail: caseycan@gmail.com.

HUMAN FACTORS
Vol. 58, No. 8, December 2016, pp. 1158–1172
DOI: 10.1177/0018720816665025

INTRODUCTION

Phishing is among the top cyberattack vectors (Symantec, 2016; Verizon Communications, 2016) threatening individuals, corporations, and critical infrastructure (Wueest, 2014). These attacks are designed to trick users into thinking an e-mail or website is legitimate and to convince them to divulge usernames and passwords or to inadvertently install malware by clicking on malicious links or attachments. Depending on the level of deception involved, it can be difficult to screen such messages automatically. As a result, human judgment plays a role in all cybersecurity systems and, by many accounts, is its weakest link (CERT, 2013; Cranor, 2008).

We use signal detection theory (SDT) methods to assess phishing vulnerability by treating phishing detection as a vigilance task (Mackworth, 1948; See, Howe, Warm, & Dember, 1995; Warm, Parasuraman, & Matthews, 2008). SDT has been used in a wide variety of contexts, including baggage screening (Wolfe, Brunelli, Rubinstein, & Horowitz, 2013), sexual intent (Farris, Treat, Viken, & McFall, 2008), medical decision making (Mohan, Rosengart, Farris, Fischhoff, & Angus, 2012), environmental risk perception (Dewitt, Fischhoff, Davis, & Broomell, 2015), and phishing detection (Kaivanto, 2014; Kumaraguru, Sheng, Acquisti, Cranor, & Hong, 2010; Mayhorn & Nyeste, 2012; Sheng, Holbrook, Kumaraguru, Cranor, & Downs, 2010; Welk et al., 2015). By quantifying performance, SDT offers metrics for analyzing system vulnerability as well as for designing and evaluating interventions to reduce it, such as training, incentives, and task restructuring (Mumpower & McClelland, 2014; Swets, Dawes, & Monahan, 2000). Such research meets a growing need to integrate human decision making and perceptual ability into cybersecurity systems (Boyce et al., 2011; Proctor & Chen, 2015).

The premise of SDT is the need to separate users' *sensitivity* or *d*′ (i.e., their ability to tell whether an e-mail is phishing) from their *response bias* or *c* (i.e., their tendency to treat an e-mail as phishing) (Macmillan & Creelman, 2004). Accuracy measures, such as the number or proportion of successful phishing attacks, are incomplete because they ignore other objectives, such as opening legitimate e-mails promptly. SDT accommodates the inevitable trade-off between hit rates (*H*, correctly identifying a signal) and false-alarm rates (*FA*, incorrectly identifying noise as signals).

The present study demonstrates a procedure for estimating individual users' sensitivity and response bias for phishing, in examining performance on two interrelated tasks: (a) *detection*, deciding whether an e-mail is legitimate, and (b) *behavior*, deciding what to do with an e-mail. Unlike many signal detection tasks, where the contingent behavior is straightforward (e.g., rescreening detected bags entails minimal costs for false positives; Wolfe et al., 2007), with phishing, detection and behavior decisions are not uniquely coupled. For example, not falling for a phishing e-mail might reflect discrimination or disinterest. As a result, we study detection and behavior separately in order to assess their respective contributions to vulnerability.

Because behavior has more immediate consequences than detection, we expected greater caution with behavior (Lynn & Barrett, 2014). However, we had no reason to expect differences in sensitivity, unless the more immediate consequences of the behavior task elicit greater effort, revealing discrimination ability not tapped by detection.

Factors That Influence Signal Detection Estimates

Previous signal detection research has identified a variety of task, individual, and environmental variables that can affect performance (Ballard, 1996). Here, we study behavior as a function of participants' awareness of two such variables: (a) signal base rate (i.e., how frequently the signal appears) and (b) costs for correct and incorrect choices (Coombs, Dawes, & Tversky, 1970; Macmillan & Creelman, 2004). These variables have typically had effects consistent with rational decision making. For example, people are more likely to identify a stimulus as noise for low-base-rate events, where it is unlikely to be a signal. Conversely, people are more likely to identify a stimulus as a signal when missing a signal is more costly and a false alarm is less costly (Lynn & Barrett, 2014; Maddox, 2002; Navalpakkam, Koch, & Perona, 2009).

The base rate and costs are related to response bias in the following equation, combining Equation 6.4 in Coombs et al. (1970) and Equation 2.6 in Macmillan and Creelman (2004):

$$\frac{P(x|s)}{P(x|n)} \geq \frac{1-p}{p}\left[\frac{C_{\mathrm{FA}}+C_{\mathrm{TN}}}{C_{\mathrm{M}}+C_{\mathrm{H}}}\right] = \beta = e^{cd'}$$

The first term is the likelihood ratio of a stimulus being a signal (*s*) or noise (*n*); *p* is the base rate of the signal; the bracketed term is the *cost ratio*, incorporating the cost of false alarms (FA), true negatives (TN), misses (M), and hits (H); and β is a measure of bias related to *c* and *d′* (as seen in the final term). When the likelihood ratio is greater than β, an observer should treat the stimulus as a signal. Assuming that *d′* remains constant with changes in task, *c* should respond to changes in *p* and the cost ratio (Lynn & Barrett, 2014). We consider both task features in the study design.

Signal base rate. Due to the volume of legitimate e-mail traffic and the use of automatic screening programs, phishing e-mails typically have a low base rate (<1%; Symantec, 2016; Verizon Communications, 2016). In the context of baggage screening, Wolfe et al. (2007) describe a prevalence effect, whereby users are biased toward identifying stimuli as noise when there is a low base rate, leading to low hit and false-alarm rates. The demands of experimental research typically lead to tasks with artificially high base rates (e.g., Mohan et al., 2012) in order to keep costs down and participants engaged. Participants are, however, typically not told the base rate, leaving it unclear whether they assume a low base rate (as in their lives) or a much higher one due to the experimental context ("They wouldn't ask me to look for phishing e-mails if they weren't going to present them fairly often"). They may also infer the base rate based on their intuitions regarding whether experimental stimuli are signals or noise (Wolfe et al., 2007). Here, we examine the effects of explicitly informing participants that the phishing base rate is 50%. If participants who receive no notice infer a 50% base rate, then notification should have little effect. If they infer a lower base rate, then their c should be much higher, indicating less caution regarding attacks.

Costs. The consequences of successful phishing can vary widely across domains. The cost of failed detection could be very high, as with critical infrastructure (e.g., an electrical grid blackout), or fairly low, as with a personal laptop (e.g., an annoying virus). Often, users have little direct guidance about those consequences beyond general cautionary messages (Carpenter, Zhu, & Kolimi, 2014). They may also have limited opportunities to learn from experience, as when time separates the attack and its damage or when users provide portals to attack distant targets. Incentives may also be misaligned, as when individuals bear the costs of avoidance actions, whereas the benefits accrue to the system (e.g., Herley, 2009, discusses rational rejection of security advice).

In detection tasks without a clear payoff structure, participants typically try to maximize accuracy (Maddox, 2002), which would produce $c = 0$ (at a 50% base rate). However, phishing avoidance is an everyday task. In order to capture participants' natural cost expectations, as best we could, we did not impose a cost structure but compared c for the detection and behavior tasks, expecting less caution for the former, with its reduced costs. Within each task, we expected individual participants' c values to be correlated with their judgments of the consequences of falling for a phishing attack. For the participants notified of the 50% base rate, we assume that β equals the cost ratio. If the base rate notification condition has no effect, we can make the same assumption for the participants without the notice. If the costs of hits (correctly identifying phishing e-mails) and true negatives (correctly identifying legitimate e-mails) are minimal, then $\beta > 1$ (and hence $c > 0$) implies a cost ratio with lower costs for misses and greater costs for false alarms. Thus, participants who judge the consequences of misses to be worse should have $\beta < 1$ and a negative (or more cautious) c.

Factors That Influence Phishing Susceptibility

Individuals' performance reflects both their ability and how well they apply it. In order to disrupt that application, attackers choose cues designed to evoke heuristic thinking and reduce systematic processing. For recipients who stop to examine messages, and possess requisite knowledge or experience, potentially useful cues include the sender, embedded URLs, grammar, spelling, sense of urgency, and subject line. Studies have, indeed, shown less susceptibility among individuals who pay greater attention to message cues, invest more cognitive effort, have more knowledge and experience, and are more suspicious (Luo, Zhang, Burd, & Seazzu, 2013; Mayhorn & Nyeste, 2012; Pattinson, Jerram, Parsons, McCormac, & Butavicius, 2012; Sheng et al., 2010; Vishwanath, Herath, Chen, Wang, & Rao, 2011; Wang, Herath, Chen, Vishwanath, & Rao, 2012; Welk et al., 2015; Wright et al., 2009; Wright & Marett, 2010). Rather than manipulate message features in order to determine participants' sensitivity to them, we use naturalistic stimuli, meant to capture everyday covariation among the cues. We assess participants' overall feeling for their discrimination ability by eliciting their confidence in their judgment, expecting more confident participants to be more knowledgeable, although not perfectly calibrated (Dhamija, Tygar, & Hearst, 2006; Lichtenstein & Fischhoff, 1980; Sheng et al., 2007). We also use a measure of dispositional suspiciousness, expecting those higher on that trait to perceive worse consequences and be more cautious but not to differ in their discrimination ability.

Aim of Study

We demonstrate an approach applying SDT to phishing detection with two interrelated tasks, detection and behavior in response to phishing, and manipulating three task variables: (a) which task comes first, detection or behavior (Experiment 1); (b) whether participants perform both tasks (Experiment 1) or just one (Experiment 2); and (c) whether participants are told, or must infer, the base rate of phishing messages. For each stimulus, we measure participants' confidence, judgments of consequences, and response time.

METHOD

Sample

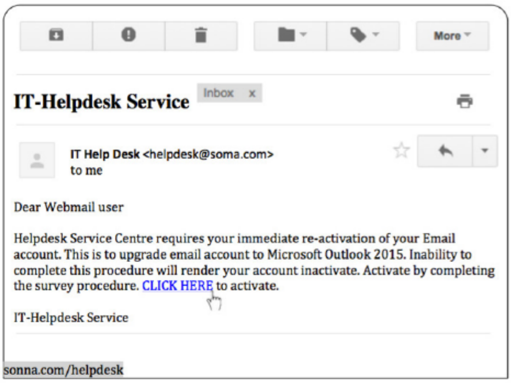
We recruited participants from U.S. Amazon Mechanical Turk (mTurk), a crowd-sourced digital marketplace often used for behavioral research (Paolacci, Chandler, & Ipeirotis, 2010). Although mTurk samples are not representative of the general U.S. population, they are more varied than convenience samples, like university students (Crump, McDonnell, & Gureckis, 2013; Mason & Suri, 2012). Often, mTurk studies recruit some participants who click through tasks without performing them or perform multiple tasks simultaneously, devoting limited attention to each (Downs, Holbrook, Sheng, & Cranor, 2010). As a result, we use attention checks to measure participants' engagement. This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at Carnegie Mellon University. Informed consent was obtained from each participant.

Design

Following the scenario-based design of Kumaraguru et al. (2010) and Pattinson et al. (2012), participants reviewed e-mails of a fictitious persona. To reduce participant burden and study costs, phishing e-mails appear at a high base rate (50%), relative to real-world settings (<1%). We randomly assigned participants to conditions created by crossing three task variables: (a) task order (Experiment 1 only), (b) task type (Experiment 2 only), and (c) notification of base rate.

Stimuli

Participants reviewed e-mails on behalf of Kelly Harmon, an employee at the fictional Soma Corporation, about whom they received a brief description. Phishing e-mails were adapted from public archives and descriptions in news articles. Each contained one or more of the following features often associated with phishing: (a) impersonal greeting, (b) suspicious URLs with a deceptive name or IP address, (c) unusual content based on the ostensible sender and subject, (d) requests for urgent action, and (e) grammatical errors or misspellings (Downs, Holbrook, & Cranor, 2006). The URL was the most valid cue for identifying a phishing e-mail. Legitimate e-mails were adapted from personal e-mails and example e-mails on the Internet, leading to some phishing cues appearing in legitimate e-mails (e.g., misspelling). Figure 1 shows a phishing e-mail. We randomized the use of personal greetings across all e-mails but did not systematically vary other cues. All stimuli mimicked the Gmail format and appear in the supplementary materials, available at http://hfs.sagepub.com/supplemental.



IT-Helpdesk Service

IT Help Desk <helpdesk@soma.com>
to me

Dear Webmail user

Helpdesk Service Centre requires your immediate re-activation of your Email account. This is to upgrade email account to Microsoft Outlook 2015. Inability to complete this procedure will render your account inactivate. Activate by completing the survey procedure. CLICK HERE to activate.

IT-Helpdesk Service

sonna.com/helpdesk

Figure 1. A phishing e-mail with all five cues.

Measures

Before viewing the stimuli, participants saw one of two messages regarding the base rate: "Approximately half of the e-mails are phishing e-mails" or "Phishing e-mails are included" (notification of base rate). In Experiment 1, participants answered the following questions for each e-mail: (a) "Is this a phishing e-mail?" (yes/no; detection), (b) "What would you do if you received this e-mail?" (with multiple-choice options from Sheng et al., 2010; behavior), (c) "How confident are you in your answer?" (50%–100%; confidence), and (d) "If this was a phishing e-mail and you fell for it, how bad would the consequences be?" (1 = *not bad at all*, 5 = *very bad*; perceived consequences). Experiment 2 randomly assigned participants to answer either Question a or b, rather than both.

To calculate *d'* and *c*, the behavior decisions were converted to binary data. Responses of *click link* and *reply*, the two actions that could expose users to negative consequences, were interpreted as indicating that participants saw the message as "legitimate"; all other responses were categorized as "phishing."

We included four attention checks. At the beginning, two multiple-choice questions asked about the task description: (1) "Where does

Kelly Harmon work?" and (2) "What is a phishing e-mail?" Embedded in the task were two e-mail stimuli used as attention checks: (3) "If you are reading this, please answer that this is a phishing e-mail" and (4) "If you are reading this, please answer that this is NOT a phishing e-mail." Many participants saw the "legitimate" stimulus check as suspicious and identified it as phishing, thereby failing the check (44 for Experiment 1 and 33 for Experiment 2). Therefore, we removed it from the analysis. Attention was measured as a binary variable based on the first three checks. Rather than removing participants who failed checks, we used attention as a predictor in the regression analyses (described later). We found similar results (see supplementary materials) when excluding the 10 participants who failed two of three additional attention checks: illogical response (e.g., clicking the link on an e-mail identified as phishing), spending less than 10 s on more than one e-mail, and $d' < 0$.

We measured the time spent on the phishing information (phish info time) and e-mails (median time/e-mail). We used gender, age, and education to measure demographic differences. (See supplementary materials for details on treatment of these variables.)

SDT Analysis

SDT assumes that both signals (phishing) and noise (legitimate e-mails) can be represented as distributions of stimuli that vary on the decision variable (here, having properties of phishing e-mails). The further apart the distributions, the greater the sensitivity or d'. The response bias, c, reflects how biased users are toward treating a stimulus as signal or noise. It is measured by how far their decision threshold is from the intersection of the two distributions. A negative response bias ($c < 0$) reflects a tendency to call uncertain stimuli signals. With phishing as the signal, negative values of c reflect a tendency to call uncertain messages phishing, indicating greater aversion to misses (treating phishing messages as legitimate) than to false alarms (treating legitimate messages as phishing).

We estimated the SDT parameters by assuming the signal and noise distributions were Gaussian with equal variance (Lynn & Barrett, 2014). To accommodate cases in which participants identified all stimuli correctly or incorrectly, producing hit (H) or false alarm (FA) rates of 0 or 1, a log-linear correction added 0.5 to the number of hits and false alarms and 1 to the number of signals (phishing e-mails) or noise (legitimate e-mails) (Hautus, 1995). Thus,

$$\mathrm{H} = (\text{hits} + 0.5)/(\text{signals} + 1)$$

$$\mathrm{FA} = (\text{false alarms} + 0.5)/(\text{noise} + 1)$$

$$d' = \mathrm{z(H)} - \mathrm{z(FA)}$$

$$c = -0.5[\mathrm{z(H)} + \mathrm{z(FA)}]$$

EXPERIMENT 1

Procedure

Participants received information about phishing and then evaluated 40 e-mails. The information was the PhishGuru comic strip from Kumaraguru et al. (2010). It noted that attackers can forge senders and warned, "Don't trust links in an e-mail." For the e-mail evaluation task, participants examined 19 legitimate e-mails, 19 phishing e-mails, and two attention-check e-mails. For each e-mail, participants performed the detection and behavior tasks, then assessed their confidence in their judgments and the perceived consequences if the e-mail was phishing. The order of the e-mails was randomized for each participant. The order of the detection and behavior tasks was randomized across participants.

Sample

Of the 162 participants who started the experiment, 152 finished. They were paid $5. According to self-reports, 58% were female and 45% had at least a bachelor's degree. The mean age was 32 years old, with a range from 19 to 59.

Of the 152 participants, 15 failed at least one attention check. For the scenario checks, three failed the work question and nine the phishing question. For the stimuli check, five failed the phishing version. They spent a minute or two (*Mdn* = 0.95 min, *M* = 3.2 min, *SD* = 11.5 min) on the phishing information and just under a minute per e-mail (*Mdn* = 43 s, *M* = 52 s, *SD* = 38 s), with a median overall time of 40 min.

TABLE 1: Signal Detection Theory Performance Parameter Estimates

	Detection Task		Behavior Task		
Variable	Experiment 1 *M* (*SD*)	Experiment 2 *M* (*SD*)	Experiment 1 *M* (*SD*)	Experiment 2 *M* (*SD*)	Typical Range
d′	0.96 (0.64)	0.98 (0.80)	0.39 (0.50)	0.41 (0.54)	0 to 4
c	0.32 (0.46)	0.30 (0.44)	−0.54 (0.66)	−0.75 (0.73)	−2 to 2
AUC	0.71 (0.12)	0.70 (0.14)	0.66 (0.12)	0.66 (0.12)	0.5 to 1
β	1.59 (1.13)	1.73 (1.49)	0.88 (0.56)	0.95 (0.43)	0 to 10
H	0.56 (0.19)	0.57 (0.19)	0.72 (0.21)	0.79 (0.16)*	0 to 1
FA	0.24 (0.16)	0.25 (0.18)	0.61 (0.21)	0.65 (0.25)	0 to 1
Accuracy	0.67 (0.11)	0.67 (0.13)	0.56 (0.08)	0.43 (0.09)***	0 to 1

Note. AUC = area under the curve; H = hit rate; FA = false alarm rate. Significant difference between Experiments 1 and 2 based on two-sided *t* test where $^{*}p < .05$, $^{***}p < .001$.

Results and Discussion

Phishing detection performance. We estimated d' and c for the detection and behavior tasks separately, denoted by subscripts D and B, respectively. Table 1 shows aggregate performance. Figure 2 shows individual performance. Additional analysis found that d' and c were constant over the course of the experiment (i.e., no learning occurred; see supplementary materials for details). We also estimated the area under the curve (AUC) for the individual receiver operating characteristic (ROC) curves, which is comparable to d', and β, which is a function of d' and c. Closer inspection (detailed in the supplementary materials) suggests that some participants in both tasks appeared to treat misses and false alarms as equally costly ($\beta = 1$), effectively making accuracy their criterion. In the behavior task, most participants appeared to minimize misses ($\beta_B < 1$). However, their thresholds varied widely for the detection task, with most aiming to minimize false alarms ($\beta_D > 1$). As expected, average perceived consequences was negatively correlated with both β_D, $r(150) = -.26$, $p = .001$, and β_B, $r(150) = -.25$, $p = .002$, indicating that participants who perceived worse consequences had lower implicit cost ratios. Participants with a higher β_D also had higher β_B, $r(150) = .36$, $p < .001$.

Detection task. Participants' mean sensitivity ($d'_D = 0.96$) indicated modest detection ability. Their mean response bias ($c_D = 0.32$) meant that they had to be somewhat suspicious before treating a message as phishing. These parameters are equivalent to a miss rate of 44% and a false alarm rate of 24%—both of which would be punishingly high for many computer systems. As seen in Figure 2a, both parameters varied considerably across participants. Some had $d'_D < 0$, meaning they consistently misidentified stimuli. Most had positive c_D values. Such variability suggests that a system's vulnerability might be very different depending on whether it was determined primarily by the average user, the worst user (in terms of d' or c), or the best user (as a sentinel for problems).

Behavior task. When asked how they would respond to each e-mail, participants demonstrated lower sensitivity ($d' = 0.39$), along with a bias toward not clicking on links ($c = -0.54$). This combination is equivalent to a miss rate of 28% and a false alarm rate of 61%, also punishingly high for many systems. Figure 2b shows the variability in individual performance. Performance on the two tasks was correlated. Participants with a high d' in the detection task tended to also have a higher d' for the behavior task, $r(150) = .61$, $p < .001$. The same was true for response bias, $r(150) = .66$, $p < .001$.

Figure 3a shows responses on the behavior task, based on whether the participant judged a message to be phishing or legitimate in the detection task. Although participants sometimes acted cautiously with messages that they perceived as legitimate (e.g., checking the link or sender), they rarely chose to "click link or open

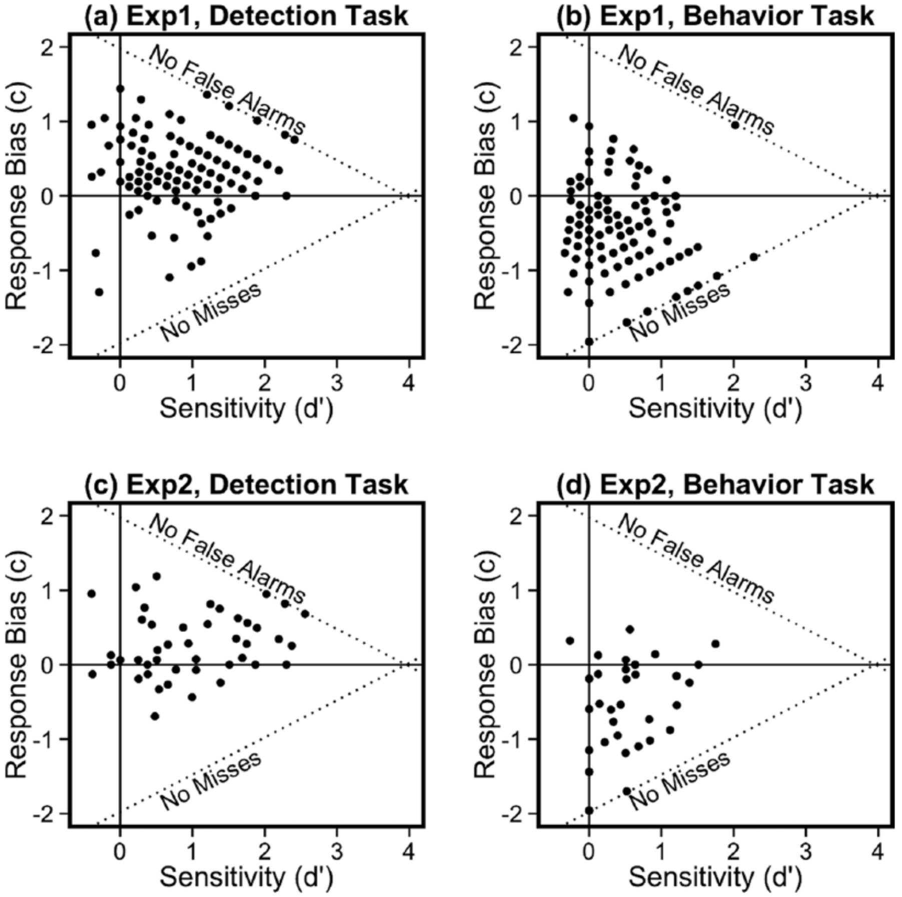


Figure 2. Individual variation for detection and behavior tasks in Experiments 1 and 2. The dotted lines denote the mathematical bounds for performance with a false alarm or miss rate of 0%.

attachment" for e-mails they perceived as phishing. Figure 3b shows these actions as a function of whether the messages were actually legitimate or phishing. Given participants' imperfect detection ability, behaviors consistent with their beliefs sometimes led to inappropriate actions. Thus, despite the bias toward not clicking on links revealed in c_B, participants still succumbed to many phishing attacks. They knew what to do with legitimate and phishing e-mails, just not which they were facing.

Regression analysis. Tables 2 and 3 show multivariate linear regression models predicting individual participants' d' and c between subjects. Model 1 considers the two between-subjects experimental task variables: (a) task order and (b) notification of base rate. Model 2 adds participants' other responses: attention, phishing information time, median time per e-mail, mean confidence, and mean perceived consequences. Model 3 adds the three demographic measures: age, gender, and college degree. Given the number of statistical tests (11), we use alpha = .01 as the threshold for significance and include tests at the alpha = .05 level, for the reader's convenience.

Model 1: Manipulated between-subject variables. Whether participants performed the detection or the behavior task first did not predict d' or c for either task, nor did whether they received explicit notification of the base rate, $p > .01$.

Model 2: Responses to stimuli. Participants who failed the attention checks had lower sensitivity on the detection task but were no different

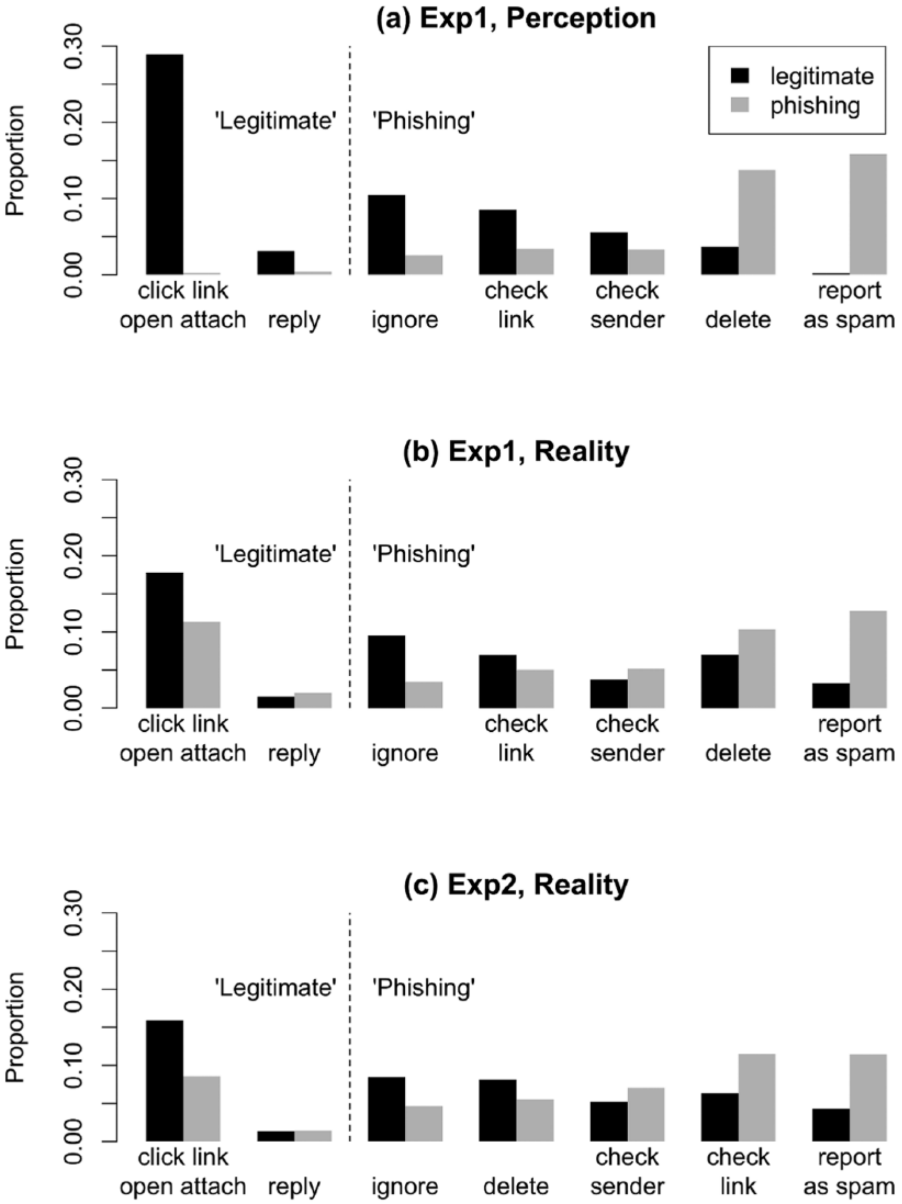


Figure 3. Proportion of behavior based on (a) perceived and (b, c) actual type of e-mail.

on the other performance parameters. Thus, users who paid less attention also exhibited lower discrimination ability but did not differ in how cautiously they acted, given their perceptions. Time spent on the phishing information was not correlated with d' or c, for either task. Participants who spent more time per e-mail were less likely to click on links (i.e., lower c_B), but were no different on the other parameters. The median time spent on each e-mail was uncorrelated to confidence and perceived consequences, $p > .01$.

For the detection task, participants' sensitivity was positively correlated with their confidence, consistent with having some metacognitive ability (i.e., knowing how much they know). Participants who were more likely to treat e-mails as legitimate (i.e., higher c_D) also tended

TABLE 2: Regression Models for *d*′ (Experiment 1)

	Detection Task			Behavior Task		
	Model 1: Task Manipulations	Model 2: Stimuli Variables	Model 3: Individual Variables	Model 1: Task Manipulations	Model 2: Stimuli Variables	Model 3: Individual Variables
Variable	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)
Intercept	0.91 (0.09)***	−2.12 (0.67)**	−1.32 (0.98)	0.34 (0.07)***	−1.28 (0.56)*	−0.09 (0.83)
Knowledge of base rate	0.07 (0.10)	0.04 (0.10)	0.02 (0.10)	0.11 (0.08)	0.10 (0.08)	0.10 (0.08)
Task order (detection = 1)	0.02 (0.10)	0.08 (0.10)	0.04 (0.10)	0 (0.08)	−0.01 (0.08)	−0.05 (0.09)
Attention (pass = 1)		0.52 (0.18)**	0.49 (0.18)**		0.15 (0.15)	0.12 (0.15)
Log(phish info time)		0.05 (0.04)	0.05 (0.04)		−0.03 (0.04)	−0.03 (0.03)
Median time/e-mail		0.40 (0.22)	0.48 (0.23)*		0.04 (0.18)	0.17 (0.19)
Average confidence		2.45 (0.65)***	2.23 (0.67)**		1.34 (0.55)*	1.11 (0.57)
Average perceived consequences		0.07 (0.08)	0.08 (0.08)		0.09 (0.06)	0.11 (0.06)
Log(age)			−0.22 (0.21)			−0.33 (0.17)
Gender (male = 1)			0.11 (0.10)			0.06 (0.09)
College (college degree = 1)			0.19 (0.10)			0.10 (0.09)
N	152	142	142	152	142	142
Adjusted R^2	−0.01	0.15	0.16	0	0.03	0.05
F	0.24	4.43***	3.71***	0.84	1.59	1.68

Note. Confidence was evaluated from 0.5 to 1 and perceived consequences were evaluated from 1 to 5.
*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 3: Regression Models for *c* (Experiment 1)

	Detection Task			Behavior Task		
	Model 1: Task Manipulations	Model 2: Stimuli Variables	Model 3: Individual Variables	Model 1: Task Manipulations	Model 2: Stimuli Variables	Model 3: Individual Variables
Variable	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)
Intercept	0.32 (0.07)***	−0.43 (0.47)	0.06 (0.70)	−0.67 (0.10)***	−0.64 (0.59)	0.10 (0.87)
Knowledge of base rate	0.03 (0.07)	0.01 (0.07)	0.01 (0.07)	0.16 (0.11)	0.12 (0.09)	0.13 (0.09)
Task order (detection = 1)	−0.04 (0.07)	−0.01 (0.07)	−0.01 (0.07)	0.12 (0.11)	0.11 (0.09)	0.11 (0.09)
Attention (pass = 1)		0.08 (0.13)	0.08 (0.13)		−0.19 (0.16)	−0.19 (0.16)
Log(phish info time)		0.01 (0.03)	0.01 (0.03)		0 (0.04)	0.01 (0.04)
Median time/e-mail		0.03 (0.15)	0.10 (0.16)		−0.79 (0.19)***	−0.70 (0.20)***
Average confidence		1.68 (0.46)***	1.81 (0.48)***		2.34 (0.58)***	2.38 (0.59)***
Average perceived consequences		−0.24 (0.05)***	−0.24 (0.05)***		−0.43 (0.07)***	−0.42 (0.07)***
Log(age)			−0.17 (0.15)			−0.22 (0.18)
Gender (male = 1)			−0.13 (0.07)			−0.14 (0.09)
College (college degree = 1)			0.02 (0.07)			−0.13 (0.09)
N	152	142	142	152	142	142
Adjusted R^2	0	0.18	0.18	0.01	0.37	0.39
F	0.25	5.34***	4.16***	1.57	13.02***	9.85***

Note. Confidence was evaluated from 0.5 to 1 and perceived consequences were evaluated from 1 to 5.
*$p < .05$. **$p < .01$. ***$p < .001$.

to be more confident. Participants who saw more severe consequences were less likely to identify e-mails as legitimate but had no difference in sensitivity. For the behavior task, participants who were more likely to click on links (i.e., higher c_B) tended to be more confident and perceive fewer consequences. We observed no differences in terms of sensitivity.

Model 3: Demographics. No demographic variable was a significant predictor of d' or c, for either task, $p > .01$.

For both tasks, d' and c were unrelated to whether participants were notified of the base rate or which task they completed first. Notification may have had no effect because participants who received no notice assumed a base rate close to 50% (because it was an experiment) or because those who received notice did not (or could not) incorporate the stated base rate in their responses given that there was no feedback (Goodie & Fantino, 1999; Newell & Rakow, 2007). Task order might have had no effect because once participants performed both tasks on a few stimuli, the two merged in their minds. Experiment 2 examines this possibility, as well as replicating the study as a whole, by having each participant perform just one task.

EXPERIMENT 2

Procedure

Experiment 2 repeats the procedure of Experiment 1, except that participants were randomly assigned to perform either the detection or the behavior task.

Sample

One hundred participants completed the online experiment, with 52 performing the detection task and 48 the behavior task. Participants who had completed Experiment 1 were not eligible for Experiment 2 (and were screened using mTurk qualifications). They were paid $5. The median time spent was 30 min. According to self-reports, 48% were female and 40% had at least a bachelor's degree. The mean age was 33 years old, with a range of 19 to 60.

Of the 100 participants, nine failed at least one attention check. For the scenario checks, one participant failed the work question and four the phishing question. Four failed the stimulus check. Presumably because participants completed only one task, the median time per e-mail was shorter (Mdn = 29 s, M = 43 s, SD = 49 s), $t(183) = 2.87, p = .005$. There was no significant difference in time spent on the phishing information, $p > .05$.

Results and Discussion

In Experiment 2, participants explicitly performed only one of the two tasks. As seen in Table 1 and Figure 2, performance was remarkably similar to Experiment 1, where participants performed both. Two-sided t tests showed no significant differences ($p > .05$) between the studies in sensitivity, response bias, confidence, or perceived consequences. The supplementary materials provide additional detail.

One possible explanation for the similarity of the results in the two experiments is that people implicitly make a detection decision when making a behavioral choice and vice versa. As a result, the second task is there implicitly, even when not performed explicitly. If so, then the similarity of the results suggests the robustness of performance on these tasks, which was also unaffected by the order in which they were performed and whether the base rate was stated. The few differences between the experiments, reported in the supplementary materials, were in whether coefficients in the regressions were above or below statistical significance (with the signs being consistent).

GENERAL DISCUSSION

SDT disentangles and quantifies sensitivity and response bias. Here, we apply it to distinguishing phishing e-mails from legitimate ones, looking separately at detection (is this message phishing?) and behavior (how will you respond to it?), building on previous research (Kumaraguru et al., 2010; Pattinson et al., 2012; Sheng et al., 2010; Vishwanath et al., 2011; Wright & Marett, 2010). After reviewing phishing information, participants evaluated 40 e-mail messages on behalf of a fictitious recipient. For each message, they expressed their confidence in their evaluation and rated the severity of the consequences if the e-mail was phishing.

Experimental manipulations varied whether the detection and behavior tasks were performed together or separately, which was done first (when together), and whether the 50% base rate of phishing messages was stated explicitly.

Our results suggest four primary findings. First, participants' behavior almost always reflected appropriate or cautious actions, given their detection beliefs (Figure 3). However, their imperfect detection ability meant that such conditionally appropriate behavior still allowed many successful phishing attacks. Thus, it appears that users have learned what to do about phishing but not when to do it.

Second, the two tasks, deciding whether a message is legitimate and what to do about it, are naturally intertwined. In Experiment 1, performance on the two tasks was correlated, such that participants who had a higher *d*′ for one also had higher *d*′ for the other. Moreover, performance was the same, whichever task was completed first, suggesting that the two could not be separated. Experiment 2 showed similar performance with participants who explicitly performed just one of the tasks. Given how intertwined the two tasks seem to be, interventions that address one might naturally address the other. An intervention that succeeded in separating them might improve detection, by focusing users on that task before moving on to behavior, and improve behavior, by allowing time to reflect on the limits to their detection ability. However, as Herley (2009, 2014) observed, slowing the process degrades the user experience, hence might be rejected, even if that is just what users need.

Third, the differences between c_D and c_B suggest that participants used different decision strategies for the two tasks. SDT research has found that participants' response bias (*c*) is sensitive to both the base rate and the costs of correct and incorrect choices. The present results suggest that all participants assumed roughly the same (50%) base rate. Stating that rate explicitly made no difference in either experiment, nor was there evidence of learning over the course of the experiment. Therefore, differences in *c* can be attributed to differences in perceived costs. Although the experiment imposed no actual costs, participants might reasonably have imported cost expectations from their everyday lives.

Responses to the detection task indicated that most participants treated false alarms as more costly than misses ($\beta > 1$), whereas the ratio was reversed for the behavior tasks ($\beta < 1$). Wickelgren (1977) shows how, even when payoffs are clear, people may lack the feedback needed to estimate how well they are achieving their desired trade-offs. Thus, our estimates of response bias represent the trade-offs that participants achieved and not necessarily those that they intended. To the extent that these estimates capture participants' actual preferences, they suggest users engage in relatively lax screening for detection, in contrast to more rigorous evaluation for behavior.

Fourth, individual performance varies widely, for both *d*′ and *c*. In the regression analyses, the most consistent predictors were participants' confidence in their ability and perception of the consequences. Confidence was strongly related to *d*′ for the detection task and more weakly for the behavior task—consistent with the common result that confidence is positively, but imperfectly, correlated with knowledge (Fischhoff & MacGregor, 1986; Lichtenstein & Fischhoff, 1980; Moore & Healy, 2008; Parker & Stone, 2014). For both tasks, more confident individuals had higher values of *c* and hence were more willing to treat messages as legitimate. Participants who saw greater consequences had lower values of *c* and hence were less willing to treat messages as legitimate, a result found in other studies of phishing detection (Sheng et al., 2010; Welk et al., 2015; Wright & Marett, 2010). In future research, better measurement of perceived consequences might improve these predictions and clarify the causal relationship.

Future research using SDT also offers the possibility of assessing the effects of interventions that might affect both *d*′ and *c*, such as brief training exercises at a high base rate with full feedback (Kaivanto, 2014; Wolfe et al., 2007, 2013), phishing detection games (Kumaraguru et al., 2010; Sheng et al., 2010; Welk et al., 2015), and communicating cost information (Davinson & Sillence, 2010; Hardee, Mayhorn, & West, 2006). Authors of that research could also examine the effects of targeting users who pose the greatest threat to system performance (Egelman & Peer, 2015), such as those identified here with $d' < 0$—indicating no detection ability or even systematic confusion.

The patterns observed in these two experiments were robust across three manipulations that could, plausibly, have affected them, namely, notifying participants of the base rate, separating the detection and behavior tasks, and varying their order. Although that robustness increases confidence in these patterns, we would hesitate to generalize the performance estimates observed here beyond the present experimental setting. Speculatively, sensitivity might be better or worse with individuals' personal e-mails, found in a more familiar context but also amid the distractions of everyday life, where monitoring phishing is a secondary task. Indeed, performance here might be a best-case scenario, with phishing the primary task and a high base rate of signals (Wolfe et al., 2007). Nonetheless, performance here was still imperfect, despite evidence suggesting that participants were trying (e.g., attention checks, orderly regression results, robustness of replication, and differential responses to the detection and behavior tasks that plausibly reflect real-world sensitivity).

Overall, participants exhibited cautious, informed behavior. However, their detection ability was sufficiently poor that their behavior could imperil computer systems dependent on this human element. Based on these results, two promising places for system operators to focus are helping users to understand the consequences of successful phishing attacks and the validity of the signal sent by their own feelings of confidence.

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, and Lorrie Cranor for comments on study concept. In addition, we thank Carnegie Mellon's Behavior, Decision, and Policy Working Group and the CyLab Usable Privacy and Security Laboratory for their feedback.

KEY POINTS

- Users had imperfect ability to determine whether e-mail messages were legitimate or phishing.
- Users knew how to deal with phishing attempts but not always when to execute those actions, given their limited detection ability.
- Interventions could focus on helping users to understand the consequences of falling for phishing attacks and how much to trust their ability to detect them.

SUPPLEMENTARY MATERIALS

The online supplementary material is available at http://hfs.sagepub.com/supplemental.

REFERENCES

Ballard, J. C. (1996). Computerized assessment of sustained attention: A review of factors affecting vigilance performance. *Journal of Clinical and Experimental Neuropsychology*, *18*(6), 843–863. doi:10.1080/01688639608408307

Boyce, M. W., Duma, K. M., Hettinger, L. J., Malone, T. B., Wilson, D. P., & Lockett-Reynolds, J. (2011). Human performance in cybersecurity: A research agenda. In *Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting* (pp. 1115–1119). Santa Monica, CA: Human Factors and Ergonomics Society. http://doi.org/10.1177/1071181311551233

Carpenter, S., Zhu, F., & Kolimi, S. (2014). Reducing online identity disclosure using warnings. *Applied Ergonomics*, *45*(5), 1337–1342. http://doi.org/10.1016/j.apergo.2013.10.005

CERT, Insider Threat Team. (2013). *Unintentional insider threats: A foundational study* (CMU/SEI-2013-TN-022). Retrieved from http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=58744

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice Hall.

Cranor, L. F. (2008). A framework for reasoning about the human in the loop. In *Proceedings of the First Conference on Usability, Psychology and Security* (Article 1). Berkeley, CA: USENIX Association.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE*, *8*(3), 1–18.

Davinson, N., & Sillence, E. (2010). It won't happen to me: Promoting secure behaviour among Internet users. *Computers in Human Behavior*, *26*(6), 1739–1747. http://doi.org/10.1016/j.chb.2010.06.023

Dewitt, B., Fischhoff, B., Davis, A., & Broomell, S. B. (2015). Environmental risk perception from visual cues: The psychophysics of tornado risk perception. *Environmental Research Letters*, *10*(12), 1–15. http://doi.org/10.1088/1748-9326/10/12/124009

Dhamija, R., Tygar, J. D., & Hearst, M. (2006). Why phishing works. In *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 581–590). New York, NY: ACM.

Downs, J., Holbrook, M. B., & Cranor, L. F. (2006). Decision strategies and susceptibility to phishing. In *Proceedings of Second*

Symposium on Usable Privacy and Security (pp. 79–90). New York, NY: ACM.

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system? Screening Mechanical Turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2399–2402). New York, NY: ACM.

Egelman, S., & Peer, E. (2015). The myth of the average user. In *Proceedings of the 2015 New Security Paradigms Workshop* (pp. 16–28). New York, NY: ACM.

Farris, C., Treat, T. A., Viken, R. J., & McFall, R. M. (2008). Perceptual mechanisms that characterize gender differences in decoding women's sexual intent. *Psychological Science*, *19*(4), 348–354.

Fischhoff, B., & MacGregor, D. (1986). Calibrating databases. *Journal of American Society for Information Sciences*, *37*(4), 222–233.

Goodie, A. S., & Fantino, E. (1999). What does and does not alleviate base-rate neglect under direct experience. *Journal of Behavioral Decision Making*, *12*, 302–335.

Hardee, J. B., Mayhorn, C. B., & West, R. (2006). I downloaded what? An examination of computer security decisions. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (pp. 1817–1820). Santa Monica, CA: Human Factors and Ergonomics Society.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of *d*'. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51.

Herley, C. (2009). So long, and no thanks for the externalities: The rational rejection of security advice by users. In *Proceedings of the New Security Paradigms Workshop* (pp. 1–12). New York, NY: ACM.

Herley, C. (2014). More is not the answer. *IEEE Security & Privacy*, *12*, 14–19.

Kaivanto, K. (2014). The effect of decentralized behavioral decision making on system-level risk. *Risk Analysis*, *34*(12), 2121–2142. http://doi.org/10.1111/risa.12219

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.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, *26*, 149–171.

Luo, X. R., Zhang, W., Burd, S., & Seazzu, A. (2013). Investigating phishing victimization with the Heuristic-Systemic Model: A theoretical framework and an exploration. *Computers & Security*, *38*(C), 28–38.

Lynn, S. K., & Barrett, L. F. (2014). "Utilizing" signal detection theory. *Psychological Science*, *25*(9), 1663–1673. doi:10.1177/0956797614541991

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*(1), 6–21. doi:10.1080/17470214808416738

Macmillan, N. A., & Creelman, D. C. (2004). *Detection theory: A user's guide*. New York, NY: Psychology Press.

Maddox, W. T. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior*, *78*, 567–595.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.

Mayhorn, C. B., & Nyeste, P. G. (2012). Training users to counteract phishing. *Work*, *41*, 3549–3552. http://doi.org/10.3233/WOR-2012-1054-3549

Mohan, D., Rosengart, M. R., Farris, C., Fischhoff, B., & Angus, D. C. (2012). Sources of non-compliance with clinical practice guidelines in trauma triage: A decision science study. *Implementation Science*, *7*(103), 1–10.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

Mumpower, J. L., & McClelland, G. H. (2014). A signal detection theory analysis of racial and ethnic disproportionality in the referral and substantiation processes of the U.S. child welfare services system. *Judgment and Decision Making*, *9*(2), 114–128.

Navalpakkam, V., Koch, C., & Perona, P. (2009). Homo economicus in visual search. *Journal of Vision*, *9*(1), 1–16.

Newell, B. R., & Rakow, T. (2007). The role of experience in decisions from description. *Psychonomic Bulletin & Review*, *14*(6), 1133–1139.

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

Parker, A. M, & Stone, E. R. (2014). Identifying the effects of unjustified confidence versus overconfidence. *Journal of Behavioral Decision Making*, *27*, 134–145.

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.

Proctor, R. W., & Chen, J. (2015). The role of human factors/ergonomics in the science of security: Decision making and action selection in cyberspace. *Human Factors*, *57*(5), 721–727. http://doi.org/10.1177/0018720815585906

See, J. E., Howe, S. R., Warm, J. S., & Dember, W. N. (1995). Meta-analysis of the sensitivity decrement in vigilance. *Psychological Bulletin*, *117*(2), 230–249.

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* (pp. 373–382). New York, NY: ACM.

Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L. F., Hong, J., & Nunge, E. (2007). Anti-phishing Phil: The design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the Third Symposium on Usable Privacy and Security* (pp. 88–99). New York, NY: ACM.

Swets, J. A., Dawes, R., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*(1), 1–26.

Symantec. (2016). *Internet security threat report*. Retrieved from https://www.symantec.com/security-center/threat-report.

Verizon Communications. (2016). *2016 data breach investigations report*. Retrieved from http://www.verizonenterprise.com/verizon-insights-lab/dbir/2016/

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:10.1016/j.dss.2011.03.002

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. http://doi.org/10.1109/TPC.2012.2208392

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors*, *50*(3), 433–441. doi:10.1518/001872008X312152

Welk, A. K., Hong, K. W., Zielinska, O. A., Tembe, R., Murphy-Hill, E., & Mayhorn, C. B. (2015). Will the "phisher-men" reel you in? *International Journal of Cyber Behavior, Psychology and Learning*, *5*(4), 1–17. http://doi.org/10.4018/IJCBPL.2015100101

Wickelgren, W. (1977). Speed–accuracy trade-off and information processing dynamics. *Acta Psychologica*, *41*, 67–85.

Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz, T. S. (2013). Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. *Journal of Vision*, *13*(3), 1–9. doi:10.1167/13.3.33

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. http://doi.org/10.1037/0096-3445.136.4.623

Wright, R., Chakraborty, S., Basoglu, A., & Marett, K. (2009). Where did they go right? Understanding the deception in phishing communications. *Group Decision and Negotiation*, *19*(4), 391–416.

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. doi:10.2753/MIS0742-1222270111

Wueest, C. (2014). *Targeted attacks against the energy sector*. Mountain View, CA: Symantec. Retrieved from http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/targeted_attacks_against_the_energy_sector.pdf

Casey Inez Canfield is a PhD candidate in the Department of Engineering and Public Policy, Carnegie Mellon University. She received a BS in engineering: systems from Olin College of Engineering in 2010.

Baruch Fischhoff is the Howard Heinz University Professor in the Departments of Engineering and Public Policy and of Social and Decision Science, Carnegie Mellon University. He received his PhD in psychology from the Hebrew University of Jerusalem in 1975.

Alex Davis is an assistant professor in the Department of Engineering and Public Policy, Carnegie Mellon University. He received his PhD in behavioral decision research from Carnegie Mellon University in 2012.

Date received: December 30, 2015
Date accepted: July 23, 2016