Topic: Cognition

Quantifying Phishing Susceptibility for Detection and Behavior Decisions

Casey Inez Canfield ¹

Baruch Fischhoff ^{1, 2}

Alex Davis ¹

¹ Department of Engineering and Public Policy,
Carnegie Mellon University (Pittsburgh, PA, USA)

² Department of Social and Decision Science,
Carnegie Mellon University (Pittsburgh, PA, USA)

Published in *Human Factors* (doi: 10.1177/0018720816665025)

Acknowledgements

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 (CUPS) for their feedback.

SIGNAL DETECTION FOR PHISHING DETECTION

23 Correspondence concerning this article should be addressed to Casey Canfield,
24 Engineering and Public Policy, Carnegie Mellon University, 5000 Forbes Ave,
25 Pittsburgh, PA 15213. E-mail: caseycan@cmu.edu.

Abstract

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 email is phishing; and *behavior*, deciding what to do with an email. 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 participants exhibiting cautious behavior, their limited detection ability left them vulnerable to phishing attacks. Greater sensitivity was positively correlated with confidence. Greater willingness to treat emails 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

SIGNAL DETECTION FOR PHISHING DETECTION

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

Précis: We use signal detection theory to measure vulnerability to phishing attacks for two interrelated tasks: *detection*, deciding whether an email is phishing, and *behavior*, deciding what to do with an email. Although participants exhibited cautious behavior, they still succumbed to phishing attacks due to their limited detection ability.

59 Quantifying Phishing Susceptibility for Detection and Behavior Decisions

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

68 We use signal detection theory (SDT) methods to assess phishing vulnerability by
69 treating phishing detection as a vigilance task (Mackworth, 1948; See et al., 1995; Warm
70 et al., 2008). SDT has been used in a wide variety of contexts, including baggage
71 screening (Wolfe et al., 2013), sexual intent (Farris et al., 2008), medical decision-
72 making (Mohan et al., 2012), environmental risk perception (Dewitt et al., 2015), and
73 phishing detection (Kaivanto, 2014; Kumaraguru et al., 2010; Mayhorn & Nyeste, 2012;
74 Sheng et al., 2010; Welk et al., 2015). By quantifying performance, SDT offers metrics
75 for analyzing system vulnerability, as well as for designing and evaluating interventions
76 to reduce it, such as training, incentives, and task restructuring (Mumpower &
77 McClelland, 2014; Swets et al., 2000). Such research meets a growing need to integrate
78 human decision-making and perceptual ability into cybersecurity systems (Boyce, 2011;
79 Proctor & Chen, 2015).

80 The premise of SDT is the need to separate users' *sensitivity* or d' (i.e., their
81 ability to tell whether an email is phishing) from their *response bias* or c (i.e., their

tendency to treat an email 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 emails promptly. SDT accommodates the inevitable tradeoff 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 email is legitimate and (b) *behavior*, deciding what to do with an email. 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 email 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 et al., 2009).

The base rate and costs are related to response bias in the following equation, combining Eq 6.4 in Coombs, Dawes & Tversky (1970) and Eq. 2.6 in MacMillan & Creelman (2004):

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

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 email traffic and the use of automatic screening programs, phishing emails typically have a low base rate (< 1%) (Symantec, 2016; Verizon, 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 emails, 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, while 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 emails) and true negatives (correctly identifying legitimate emails) 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, found 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 et al., 2013; Mayhorn & Nyeste, 2012; Pattinson et al., 2012; Sheng et al., 2010; Vishwanath et al., 2011; Wang et al., 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 co-variation 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 et al., 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 (1) demonstrate an approach applying SDT to phishing detection, (2) with two interrelated tasks, detection and behavior in response to phishing; and (3) 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' (a) confidence, (b) judgments of consequences, and (c) 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). mTurk studies often recruit some participants who click through tasks without performing them or perform multiple tasks simultaneously, devoting limited attention to each (Downs et al., 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 emails of a fictitious persona. To reduce participant burden and study costs, phishing emails 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: (1) task order (Experiment 1 only), (2) task type (Experiment 2 only), and (3) notification of base rate.

Stimuli

Participants reviewed emails on behalf of Kelly Harmon, an employee at the fictional Soma Corporation, about whom they received a brief description. Phishing emails were adapted from public archives and descriptions in news articles. Each contained one or more of the following features often associated with phishing: (1) impersonal greeting, (2) suspicious URLs with a deceptive name or IP address, (3) unusual content based on the ostensible sender and subject, (4) requests for urgent action, and (5) grammatical errors or misspellings (Downs et al., 2006). The URL was the most valid cue for identifying a phishing email. Legitimate emails were adapted from personal emails and example emails on the Internet, leading to some phishing cues appearing in legitimate emails (e.g., misspelling). Figure 1 shows a phishing email. We randomized the use of personal greetings across all emails, but did not systematically vary other cues. All stimuli mimicked the Gmail format, and appear in the Supplementary Materials.

Figure 1. A phishing email with all 5 cues.

Measures

Before viewing the stimuli, participants saw one of two messages regarding the base rate: (1) “Approximately half of the emails are phishing emails” or (2) “Phishing emails are included” (*notification of base rate*). In Experiment 1, participants answered the following questions for each email: (1) “Is this a phishing email?” (Yes/No) (*detection*); (2) “What would you do if you received this email?”, with multiple-choice options from Sheng et al. (2010) (*behavior*); (3) “How confident are you in your answer?” (50-100%) (*confidence*); and (4) “If this was a phishing email 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 1 or 2, 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 4 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 email?” Embedded in the task were two email stimuli used as attention checks: (3) “If you are reading this, please answer that this is a phishing email” and (4) “If you are reading this, please answer that this is NOT a phishing email.” 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 3 checks. Rather than removing participants who failed checks, we used attention as a predictor in the regression analyses (below). We found similar results (see Supplementary Material) when excluding the 10 participants who failed two of three additional attention checks: illogical response (e.g., clicking the link on an email identified as phishing), spending less than 10 seconds on more than one email and $d' < 0$.

We measured the time spent on the phishing information (*phish info time*) and emails (*median time/email*). We used gender, age and education to measure demographic differences. (See Supplementary Materials for details on treatment of these variables.)

Signal Detection Theory Analysis

SDT assumes that both signals (phishing) and noise (legitimate emails) can be represented as distributions of stimuli that vary on the decision variable (here, having properties of phishing emails). 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 where 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 emails) or noise (legitimate emails) (Hautus, 1995). Thus:

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

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

$$d' = z(H) - z(FA)$$

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

Experiment 1

Procedure

Participants received information about phishing and then evaluated 40 emails. 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 email.” For the email evaluation task, participants examined 19 legitimate emails, 19 phishing emails, and 2 attention check emails. For each email, participants performed the detection and behavior tasks, then assessed their confidence in their judgments and the perceived consequences

if the email was phishing. The order of the emails was randomized for each participant.

The order of the detection and behavior task 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, 3 failed the work question and 9 the phishing question. For the stimuli check, 5 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 email ($Mdn = 43$ sec, $M = 52$ sec, $SD = 38$ sec), with a median overall time of 40 minutes.

Results & 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 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 , $r(150) = -0.26$, $p = .001$, and β_B , $r(150) = -0.25$, $p = .002$, indicating that participants who perceived worse consequences had lower implicit cost ratios. Participants with a higher β_D also had higher β_B , $r(150) = 0.36$, $p < .001$.

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 email, 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, $r(150) = 0.61$, $p < .001$. The same was true for response bias, $r(150) = 0.66$, $p < .001$.

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 attachment” for emails 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 emails, just not which they were facing.

Table 1

SDT Performance Parameter Estimates

	Detection Task		Behavior Task		Typical Range
	Experiment 1 M (SD)	Experiment 2 M (SD)	Experiment 1 M (SD)	Experiment 2 M (SD)	
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: Significant difference between Experiment 1 and 2 based on two-sided t-test where

* $p < .05$, *** $p < .001$

Figure 2. Individual variation for detection and behavior tasks in Experiment 1 and 2.

The dotted lines denote the mathematical bounds for performance with a false alarm or miss rate of 0%.

SIGNAL DETECTION FOR PHISHING DETECTION

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

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: (1) task order and (2) notification of base rate. Model 2 adds participants' other responses: attention, phishing information time, median time per email, 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 's $> .01$.

Model 2: Responses to stimuli. Participants who failed the attention checks had lower sensitivity on the detection task, but were no different 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 email 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 email was uncorrelated to confidence and perceived consequences, $p > .05$.

SIGNAL DETECTION FOR PHISHING DETECTION

Table 2

Regression models for d' (Experiment 1)

	<u>Detection Task</u>			<u>Behavior Task</u>		
	Model 1: Task Manipulations B (SE)	Model 2: Stimuli Variables B (SE)	Model 3: Individual Variables B (SE)	Model 1: Task Manipulations B (SE)	Model 2: Stimuli Variables B (SE)	Model 3: Individual Variables 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/email		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

Notes: *p<.05 **p<.01 ***p<.001

Confidence was evaluated from 0.5-1 and perceived consequences were evaluated from 1-5.

SIGNAL DETECTION FOR PHISHING DETECTION

Table 3

Regression models for c (Experiment 1)

	<u>Detection Task</u>			<u>Behavior Task</u>		
	Model 1: Task Manipulations B (SE)	Model 2: Stimuli Variables B (SE)	Model 3: Individual Variables B (SE)	Model 1: Task Manipulations B (SE)	Model 2: Stimuli Variables B (SE)	Model 3: Individual Variables 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/email		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***

Notes: *p<.05 **p<.01 ***p<.001

Confidence was evaluated from 0.5-1 and perceived consequences were evaluated from 1-5.

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 emails as legitimate (i.e., higher c_D) also tended to be more confident. Participants who saw more severe consequences were less likely to identify emails as legitimate, but had no different 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 minutes. 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, 9 failed at least one attention check. For the scenario checks, 1 participant failed the work question and 4 the phishing question. Four failed the stimulus check. Presumably because participants only completed one task, the median time per email was lower ($Mdn = 29$ sec, $M = 43$ sec, $SD = 49$ sec), $t(183) = 2.87$, $p = .005$. There was no significant difference in time spent on the phishing information, $p > .05$.

Results & 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 found 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 emails 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 email 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 email 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 found 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 tradeoffs. Thus, our estimates of response bias represent the tradeoffs 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, 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 , hence were more willing to treat messages as legitimate. Participants who saw greater consequences had lower values of c , 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 between caution and confidence.

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). 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 emails, found in a more familiar context, but also amidst 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, with evidence suggesting that participants were trying: attention checks, orderly regression results, robustness of

SIGNAL DETECTION FOR PHISHING DETECTION

506 replication, and differential responses to the detection and behavior tasks that plausibly
507 reflect real-world sensitivity.

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

SIGNAL DETECTION FOR PHISHING DETECTION

513 **Key points:**

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

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. *Proceedings of Human Factors and Ergonomics Society Annual Meeting*, 55(1), 1115–1119. <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). <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, New Jersey: Prentice-Hall, Inc.
- Cranor, L. F. (2008). A Framework for Reasoning About the Human in the Loop. *Proceedings of Usability, Psychology & Security (UPSEC)*.
- 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.

- 541 Davinson, N., & Sillence, E. (2010). It won't happen to me: Promoting secure behaviour
 542 among internet users. *Computers in Human Behavior*, 26(6), 1739–1747.
 543 <http://doi.org/10.1016/j.chb.2010.06.023>
- 544 Dewitt, B., Fischhoff, B., Davis, A., & Broomell, S. B. (2015). Environmental risk
 545 perception from visual cues: the psychophysics of tornado risk perception.
 546 *Environmental Research Letters*, 10(12), 1–15. [http://doi.org/10.1088/1748-](http://doi.org/10.1088/1748-9326/10/12/124009)
 547 [9326/10/12/124009](http://doi.org/10.1088/1748-9326/10/12/124009)
- 548 Dhamija, R., Tygar, J. D., & Hearst, M. (2006). Why Phishing Works. Proceedings of
 549 *CHI*, Montreal, Quebec, Canada, 581-590.
- 550 Downs, J., Holbrook, M. B., & Cranor, L. F. (2006). Decision Strategies and
 551 Susceptibility to Phishing. Proceedings of *Symposium on Usable Privacy and*
 552 *Security (SOUPS)*.
- 553 Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are Your Participants
 554 Gaming the System? Screening Mechanical Turk Workers. Proceedings of *CHI*,
 555 Atlanta, Georgia.
- 556 Egelman, S., & Peer, E. (2015). The Myth of the Average User. Proceedings of *New*
 557 *Security Paradigms Workshop (NSPW)*, Twente, Netherlands.
- 558 Farris, C., Treat, T. A., Viken, R. J., & McFall, R. M. (2008). Perceptual Mechanisms
 559 That Characterize Gender Differences in Decoding Women's Sexual Intent.
 560 *Psychological Science*, 19(4), 348–354.
- 561 Fischhoff, B. & MacGregor, D. (1986). Calibrating databases. *Journal of American*
 562 *Society for Information Sciences*, 37, 222-233.

- 563 Goodie, A. S., & Fantino, E. (1999). What Does and Does Not Alleviate Base-Rate
564 Neglect Under Direct Experience. *Journal of Behavioral Decision Making*, 12, 302–
565 335.
- 566 Hardee, J. B., Mayhorn, C. B., & West, R. (2006). I Downloaded What? : An
567 Examination of Computer Security Decisions. Proceedings of *Human Factors and*
568 *Ergonomics Society Annual Meeting*, 1817-1820.
- 569 Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on
570 estimated values of d'. *Behavior Research Methods, Instruments, & Computers*,
571 27(1), 46–51.
- 572 Herley, C. (2009). So Long, And No Thanks for the Externalities: The Rational Rejection
573 of Security Advice by Users. Presented at *New Security Paradigms Workshop*
574 (*NSPW*), Oxford, United Kingdom.
- 575 Herley, C. (2014). More Is Not the Answer. *IEEE Security & Privacy*, 14–19.
- 576 Kaivanto, K. (2014). The Effect of Decentralized Behavioral Decision Making on
577 System-Level Risk. *Risk Analysis*, 34(12), 2121–2142.
578 <http://doi.org/10.1111/risa.12219>
- 579 Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., & Hong, J. (2010). Teaching
580 Johnny not to fall for phish. *ACM Transactions on Internet Technology*, 10(2), 1–31.
- 581 Lichtenstein, S. & Fischhoff, B. (1980). Training for calibration. *Organizational*
582 *Behavior and Human Performance*, 26, 149-171.
- 583 Luo, X. R., Zhang, W., Burd, S., & Seazzu, A. (2013). Investigating phishing
584 victimization with the Heuristic-Systemic Model: A theoretical framework and an
585 exploration. *Computers & Security*, 38(C), 28–38.

- 586 Lynn, S. K., & Barrett, L. F. (2014). "Utilizing" Signal Detection Theory. *Psychological*
 587 *Science*, 25(9), 1663-1673. doi:10.1177/0956797614541991
- 588 Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search.
 589 *Quarterly Journal of Experimental Psychology*, 1(1), 6–21.
 590 doi:10.1080/17470214808416738
- 591 Macmillan, N. A., & Creelman, D. C. (2004). *Detection Theory: A User's Guide*. New
 592 York: Psychology Press.
- 593 Maddox, W. T. (2002). Toward a Unified Theory of Decision Criterion Learning in
 594 Perceptual Categorization. *Journal of the Experimental Analysis of Behavior*, 78(3),
 595 567–595.
- 596 Mason, W. & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical
 597 Turk. *Behavior Research Methods*, 44(1), 1-23.
- 598 Mayhorn, C. B., & Nyeste, P. G. (2012). Training users to counteract phishing. *Work*, 41,
 599 3549–3552. <http://doi.org/10.3233/WOR-2012-1054-3549>
- 600 Mohan, D., Rosengart, M. R., Farris, C., Fischhoff, B., & Angus, D. C. (2012). Sources
 601 of non-compliance with clinical practice guidelines in trauma triage: a decision
 602 science study. *Implementation Science*, 7(103), 1–10.
- 603 Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological*
 604 *Review*, 115(2), 502-517.
- 605 Mumpower, J. L., & McClelland, G. H. (2014). A signal detection theory analysis of
 606 racial and ethnic disproportionality in the referral and substantiation processes of the
 607 U.S. child welfare services system. *Judgment and Decision Making*, 9(2), 114–128.

- 608 Newell, B. R., & Rakow, T. (2007). The role of experience in decisions from description.
 609 *Psychonomic Bulletin & Review*, 14(6), 1133–1139.
- 610 Navalpakkam, V., Koch, C., & Perona, P. (2009). Homo economicus in visual search.
 611 *Journal of Vision*, 9(1), 31–31.
- 612 Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon
 613 Mechanical Turk. *Judgment and Decision Making*, 5(5), 411–419.
- 614 Parker, A.M, & Stone, E.R. (2014). Identifying the effects of unjustified confidence
 615 versus overconfidence. *Journal of Behavioral Decision Making*, 27, 134.145.
- 616 Pattinson, M., Jerram, C., Parsons, K., McCormac, A., & Butavicius, M. (2012). Why do
 617 some people manage phishing e-mails better than others? *Information Management &*
 618 *Computer Security*, 20(1), 18–28.
- 619 Proctor, R. W., & Chen, J. (2015). The Role of Human Factors/Ergonomics in the
 620 Science of Security: Decision Making and Action Selection in Cyberspace.
 621 *Proceedings of Human Factors: the Journal of the Human Factors and Ergonomics*
 622 *Society*, 57(5), 721–727. <http://doi.org/10.1177/0018720815585906>
- 623 See, J. E., Howe, S. R., Warm, J. S., & Dember, W. N. (1995). Meta-Analysis of the
 624 Sensitivity Decrement in Vigilance. *Psychological Bulletin*, 117(2), 230–249.
- 625 Sheng, S., Holbrook, M. B., Kumaraguru, P., Cranor, L. F., & Downs, J. (2010). Who
 626 Falls for Phish? A Demographic Analysis of Phishing Susceptibility and
 627 Effectiveness of Interventions. *Proceedings of CHI*.
- 628 Swets, J. A., Dawes, R., & Monahan, J. (2000). Psychological Science Can Improve
 629 Diagnostic Decisions. *Psychological Science in the Public Interest*, 1(1), 1–26.

- 630 Symantec. (2016). *Internet Security Threat Report*. [https://www.symantec.com/security-](https://www.symantec.com/security-center/threat-report)
631 [center/threat-report](https://www.symantec.com/security-center/threat-report).
- 632 Verizon Communications Inc. (2016). *2016 Data Breach Investigations Report*.
633 <http://www.verizonenterprise.com/verizon-insights-lab/dbir/2016/>
- 634 Vishwanath, A., Herath, T., Chen, R., Wang, J., & Rao, H. R. (2011). Why do people get
635 phished? Testing individual differences in phishing vulnerability within an integrated,
636 information processing model. *Decision Support Systems*, 51(3), 576–586.
637 doi:10.1016/j.dss.2011.03.002
- 638 Wang, J., Herath, T., Chen, R., Vishwanath, A., & Rao, H. R. (2012). Phishing
639 Susceptibility: An Investigation Into the Processing of a Targeted Spear Phishing
640 Email. *IEEE Transactions on Professional Communication*, 55(4), 345–362.
641 <http://doi.org/10.1109/TPC.2012.2208392>
- 642 Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance Requires Hard Mental
643 Work and Is Stressful. *Human Factors*, 50(3), 433–441.
644 doi:10.1518/001872008X312152
- 645 Welk, A. K., Hong, K. W., Zielinska, O. A., Tembe, R., Murphy-Hill, E., & Mayhorn, C.
646 B. (2015). Will the “Phisher-Men” Reel You In? *International Journal of Cyber*
647 *Behavior, Psychology and Learning*, 5(4), 1–17.
648 <http://doi.org/10.4018/IJCBPL.2015100101>
- 649 Wickelgren, W. (1977). Speed-accuracy trade-off and information processing dynamics.
650 *Acta Psychologica*, 41(1), 67-85.
- 651 Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N.
652 (2007). Low target prevalence is a stubborn source of errors in visual search tasks.

- 653 *Journal of Experimental Psychology: General*, 136(4), 623–638.
- 654 <http://doi.org/10.1037/0096-3445.136.4.623>
- 655 Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz, T. S. (2013). Prevalence effects
- 656 in newly trained airport checkpoint screeners: Trained observers miss rare targets,
- 657 too. *Journal of Vision*, 13(3), 33–33. doi:10.1167/13.3.33
- 658 Wright R., Chakraborty S., Basoglu A., Marett K. (2010). Where did they go right?
- 659 Understanding the deception in phishing communications. *Group Decision and*
- 660 *Negotiation*, 19, 391–416.
- 661 Wright, R. T., & Marett, K. (2010). The Influence of Experiential and Dispositional
- 662 Factors in Phishing: An Empirical Investigation of the Deceived. *Journal of*
- 663 *Management Information Systems*, 27(1), 273–303. doi:10.2753/MIS0742-
- 664 1222270111
- 665 Wueest, C. (2014). Targeted Attacks Against the Energy Sector. Symantec.
- 666 [http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepap](http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/targeted_attacks_against_the_energy_sector.pdf)
- 667 [ers/targeted_attacks_against_the_energy_sector.pdf](http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/targeted_attacks_against_the_energy_sector.pdf).

Biographies

Casey Canfield is a PhD candidate in the Department of Engineering and Public Policy, Carnegie Mellon University. She received a B.S. 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.