Metacognition and Learning
https://doi.org/10.1007/s11409-019-09197-5

Better beware: comparing metacognition for phishing and legitimate emails

Casey Inez Canfield[1] · Baruch Fischhoff[2] · Alex Davis[2]

Received: 16 April 2018 / Accepted: 26 June 2019/Published online: 20 July 2019

Abstract
Every electronic message poses some threat of being a phishing attack. If recipients underestimate that threat, they expose themselves, and those connected to them, to identity theft, ransom, malware, or worse. If recipients overestimate that threat, then they incur needless costs, perhaps reducing their willingness and ability to respond over time. In two experiments, we examined the appropriateness of individuals' confidence in their judgments of whether email messages were legitimate or phishing, using calibration and resolution as metacognition metrics. Both experiments found that participants had reasonable calibration but poor resolution, reflecting a weak correlation between their confidence and knowledge. These patterns differed for legitimate and phishing emails, with participants being better calibrated for legitimate emails, except when expressing complete confidence in their judgments, but consistently overconfident for phishing emails. The second experiment compared performance on the laboratory task with individuals' actual vulnerability, and found that participants with better resolution were less likely to have malicious files on their home computers. That comparison raised general questions about the design of anti-phishing training and of providing feedback essential to self-regulated learning.

Keywords Phishing · Calibration · Resolution · Deception detection · Digital literacy

Introduction

Phishing attacks seek to trick recipients into believing that an email is legitimate, in order to solicit sensitive information (e.g. usernames, passwords, credit card numbers) or install malware. Spear phishing attacks use personal information (e.g. known contacts, industry language, victims' names) to create more realistic and persuasive messages. In 2018, schools and universities were the third most popular target for social engineering attacks using electronic media, after the public

✉ Casey Inez Canfield
canfieldci@mst.edu

[1] Missouri University of Science & Technology, 300 W 13th St, Rolla, MO 65409, USA

[2] Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

sector and healthcare industries (Verizon 2018). Educational institutions may be at higher risk because of the relative transparency of contact information, job roles, and names for members of their communities. At present, 96% of social engineering attacks, which include phishing and pretexting, are via email (Verizon 2018). Given the difficulty of screening many such messages automatically, human behavior plays a major role in determining the vulnerability (Boyce et al. 2011; Cranor 2008; Proctor and Chen 2015; Werlinger et al. 2009).

Unfortunately, most individuals have few opportunities to learn about phishing attacks, whose deceptive nature means that they may not receive clear, timely feedback on how well they are doing. Their personal details may be stolen without their knowledge. They may be a point of entry to a distant target (as with John Podesta and the Democratic National Committee). As a result, individuals may be miscalibrated, in the sense of not knowing how much confidence to place in their own performance (Lichtenstein and Fischhoff 1980; Lichtenstein et al. 1982; Mellers et al. 2015). Such miscalibration can expose individuals to additional risk, where individuals think that they know how to protect themselves, or take needless precautions without realizing how much they know. Educators are in a unique position to help students build good habits for navigating the communications that they receive via email, social media, and other online sources (Hodgin and Kahne 2018). The research reported here examines the quality of individuals' metacognitive assessments of their ability to detect phishing in both experimental and real-world settings.

The relationship between confidence and performance for the data described here has been reported in two previous papers (Canfield et al. 2017, 2016). The present analysis reports metacognition metrics and investigates the relationship between metacognition and individual differences as well as real-world vulnerability. This analysis makes several contributions to the literature. One is extending the study of phishing detection, from its focus on performance, to consider metacognition (also addressed in Li et al. 2016; Wang et al. 2016). The second is to compare metacognition for true and false signals (here, phishing and legitimate emails). The third is to assess the predictive validity of metacognitive performance on experimental tasks with respect to real-world outcomes. The latter comparison has implications for the design and goals of anti-phishing training.

Contextualizing phishing detection in digital literacy

There is increasing interest in helping students develop tools to avoid deceptive communications – whether fake news, false Wikipedia edits, or phishing attacks. Media literacy is the ability to find, evaluate, and create communications across platforms (e.g. paper vs. digital), and is an umbrella-concept for digital and information literacy (Koltay 2011). Most media literacy research has focused on K-12 students, but there is a need to study vulnerable adult populations as well (Lee 2018), whereas phishing detection research has focused on adults – particularly university students (e.g. Jagatic et al. 2007). This manuscript places research on anti-phishing training in the context of media and digital literacy.

Digital literacy includes the ability to search for, navigate, evaluate, synthesize, and communicate in digital platforms (Koltay 2011). Eshet-Alkalai (2004) conceives of digital literacy as the blending of multiple types of literacy including photo-visual literacy, reproduction literacy, branching literacy, information literacy, and socio-emotional literacy. The ability to detect deception, such as phishing attacks, is also a form of information literacy, a focus of library and information science research. One main component of information literacy is applying critical thinking skills to evaluate authenticity and credibility (Koltay 2011).

Metacognition and learning to detect deception

Metacognition is described generally as “cognition about cognition” (Smith et al. 2003). Here, it refers to individuals’ understanding of their ability to detect phishing emails. That task is a special case of the metacognitive ability to navigate online systems, in which limited bandwidth messages may be misleading, not just because of poor design (Blackshaw and Fischhoff 1988; Fischhoff and MacGregor 1986; Macgregor et al. 1987), but because of deliberate deception. We use probability judgments to reveal metacognitive states of mind. We score them with the metacognitive metric of the Brier score, which decomposes into calibration, resolution, and knowledge (Fleming and Lau 2014). Calibration measures the degree to which confidence judgments correspond to relative frequencies. When properly calibrated, judgments of the probability of being correct equal that actual probability. In other words, calibration requires knowing how often the items associated with each level confidence are correct. Resolution measures the variance of correct answers associated with different probability judgments. Individuals with high resolution are better able to discriminate between correct and incorrect judgments by assigning them different probabilities. Knowledge captures the predictability of the events in question. In this context, it is the percentage of correct judgments. These metrics are further defined below in the Methods section.

Over a wide variety of tasks, individuals’ confidence in their knowledge has been found to be moderately correlated with how much they actually know, indicating some metacognitive ability to distinguish correct and incorrect answers (Desender et al. 2018; Fischhoff and MacGregor 1986; Kunimoto et al. 2001; Lichtenstein et al. 1982). Calibration has also been found to be sensitive to task difficulty, with individuals tending to be overconfident with more difficult tasks and underconfident for easier ones (Lichtenstein and Fischhoff 1977), consistent with imperfect sensitivity to overall task difficulty, and to the difficulty of individual items.

Unlike most detection tasks, phishing detection involves detecting deception. Successful deception increases the difficulty of detection tasks and may reduce the reliability of cues in them. A review of deception detection studies found virtually no correlation between accuracy and confidence, confirming the general finding that people are poor lie detectors (DePaulo et al. 1997). However, the review found that participants were more confident when rating truths than lies, in studies that included such analyses. This pattern suggests that individuals have some metacognitive ability, even if they are not able to apply it in their detection decisions. A more recent meta-analysis found that, across 384 experiments, participants were able to classify 54% of truths and lies correctly and tended to exhibit a “truth bias,” in the sense of erring toward seeing uncertain items as true (Bond and Depaulo 2006). However, neither review considered metacognition for true and false messages separately.

Most research on teaching people to detect deception has occurred in the context of law enforcement. In a recent meta-analysis of 30 studies, training that focused on verbal content cues was most effective – while feedback on accuracy had small to insignificant effects (Hauch et al. 2016). Other research suggests that individuals are better at detecting deception if they know that some deception has occurred, because they are better able to interpret the available cues (Von Hippel et al. 2016). These studies suggest that any training must direct individuals’ attention to the correct cues and give feedback on whether they are using them, not simply on whether their judgment is correct. Indeed, learning science research suggests that metacognition is a critical part of the learning process. For example, students with higher metacognitive skills are more motivated to learn from their mistakes (Veenman et al. 2006; Yeung and Summerfield 2012). Thus, poor metacognitive ability may be a barrier in improving detection performance.

Metacognition and learning to detect phishing

If phishing detection follows results from research on deception detection, then there will be a weak relationship between confidence and accuracy. There may also be a "truth bias," whereby participants are more likely to accurately judge emails as legitimate than phishing. If that pattern holds, then an email judged to be legitimate is more likely to be so when an individual believes in that judgment with greater confidence (Hauch et al. 2016). Researchers have found that individuals are generally overconfident in their phishing email detection (Wang et al. 2016). In one of the few studies to assess performance of phishing and legitimate email detection separately, Kleitman et al. (2018) found that confidence was weakly, but significantly, positively correlated with accuracy for both types of emails.

Self-regulated learning research integrates results from metacognition, which focuses on cognition, and self-regulation, which focuses on human behavior, considering skills such as planning, goal-setting, and self-reflection (Dinsmore et al. 2008). Exploratory work suggests that those skills are critical to acquiring digital literacy (Greene et al. 2014). These results suggest that anti-phishing training may benefit from a focus on metacognitive outcomes, rather than just performance outcomes.

Individual and task factors influencing metacognition for deception detection

Reviews of deception detection have found no relationship between expertise and ability to detect lies. In fact, individuals with more expertise (e.g., professionals such as police officers) tend to be more overconfident and have a "lie bias" (Vrij et al. 2010). In a review of 247 studies, Bond and DePaulo (2008) found little evidence of individual differences in "receiver" (i.e. person receiving the truth or lie) detection accuracy. However, they did find individual differences in "sender" (i.e. person sending the truth or lie) ability to disguise lies.

Follow-up research suggests that individual differences in the receiver may play a more important role in online, compared to in-person, deception detection. Detection accuracy for both phishing and legitimate emails was best predicted by intelligence (as measured by the Esoteric Analogies Test) and perceived maliciousness (which was predicted by confidence) (Kleitman et al. 2018). In contrast, receiver attributes were unrelated to deception detection when conducted in-person, even for lies told in foreign languages that the receiver did not understand. Even in this context, sender attributes were the best predictor of successful deception (Law et al. 2018).

One of the few studies examining metacognitive ability in email detection found that calibration is sensitive to task-related variables (Li et al. 2016). For example, better calibrated individuals spent more time reviewing each email. Resolution was associated with both task and individual difference variables. Participants who used the internet more often, specifically for online shopping, tended to have better resolution. That study relied on self-reports of internet activity; the present research includes observed behavior related to web and downloading activity.

Research questions

In order to improve the design and targeting of behavioral interventions for conferring the metacognitive skills needed to reduce phishing vulnerability, we ask the following research questions:

1. How do calibration and resolution differ for phishing (false) and legitimate (true) emails?
2. What individual and task factors predict calibration and resolution for phishing and legitimate email?
3. How are calibration and resolution for phishing and legitimate emails related to real-world vulnerability?

Methods

The data and code for our analyses are available at https://osf.io/mhqpv/. Additional analyses of the data, focused on performance rather than metacognition, are reported elsewhere for both Experiment 1 (Canfield et al. 2016) and Experiment 2 (Canfield et al. 2017).

Sample

For Experiment 1, 152 participants were recruited from Amazon mTurk (Paolacci et al. 2010), a crowdsourcing platform. Informed consent was obtained from all participants, who were each paid \$5. According to participant reports, the mean age was 32 years old (min = 19, max = 59), 58% were female, and 45% had completed at least a 4-year college degree. The sample size for Experiment 1 was determined based on a power analysis (Canfield et al. 2016).

For Experiment 2, 98 participants were recruited from the Security Behavior Observatory (SBO), an on-going observational study of computer security behavior (Forget et al. 2014, 2016). The SBO recruited participants from the local community and consisted primarily of students and retirees. Active SBO participants were recruited for "a study about email use." Informed consent was obtained from all participants separately for our study. Their mean age was 40 years old (min = 19, max = 81), 60% were female, and 65% had completed at least a 4-year college degree. The sample size for Experiment 2 was limited by the active participants in the SBO (Canfield et al. 2017).

Pooling across both experiments, there were 250 participants, with mean age of 35 years old (SD = 14 years), 59% female, and 53% with at least a 4-year college degree. Due to the skewed distribution, a log transformation was used in the regression analyses for both experiments. SBO participants were 8 years older on average, $t(130) = -4.32$, $p < .001$, and more likely to have completed a 4-year college degree, $\chi^2(1) = 8.71$, $p = .003$. A log transformation was not used when analyzing Experiment 2 alone to be consistent with Canfield et al. (2017).

Design

The research followed the online scenario-based design of Kumaraguru et al. (2010) and Pattinson et al. (2012). Participants were introduced to a persona, "Kelly Harmon, who works at the Soma Corporation," for whom they would be checking email. They were also provided with a cartoon that served as training for how to detect phishing attacks (Kumaraguru et al. 2010). Each participant reviewed 40 emails. For each email, participants answered four questions (Canfield et al. 2016):

a) *Detection Task*: “Is this a phishing e-mail?” (yes/no)
b) *Behavior Task*: “What would you do if you received this e-mail?” (multiple choice from Click link/ Open attachment, Archive it, Reply, Delete it, Report as spam, or Other; adapted from Sheng et al. 2010)
c) *Confidence*: “How confident are you in your answer?” (50%–100% continuous scale)
d) *Perceived Consequences*: “If this was a phishing e-mail and you fell for it, how bad would the consequences be?” (Likert scale, with 1 = not bad at all and 5 = very bad)

The order of the detection and behavior tasks was randomized for each participant and remained consistent across each email that an individual participant viewed. The present analysis considers only the detection task, along with the confidence and perceived consequences measures. None of the data from the behavior task are reported here (see Canfield et al. 2016, for analyses). Previous analyses (Experiment 2 in Canfield et al. 2016) found that no systematic differences in confidence judgments elicited with just the detection task or just the behavior task.

mTurk participants sometimes click through tasks without paying attention (Downs et al. 2010). As a result, we employed four attention checks: Before reviewing any emails, participants were asked two multiple-choice questions, one each about the scenario and the task: (1) “Where does Kelly Harmon work?” and (2) “What is a phishing e-mail?” Embedded within the emails were two messages that served 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.” Unfortunately, many participants found the “legitimate” stimulus check (question 4) suspicious and identified it as phishing, thereby failing the check. As a result, we removed that attention check from the analysis, and used just the first three checks. We treated participants who answered all three correctly as paying attention. Rather than removing the 32 participants who failed attention checks, we used attention as a predictor in the regression analyses in order to assess its relationship to performance (Canfield et al. 2016).

Lastly, participants provided demographic information regarding their age, gender, and education. We also measured the time spent on the phishing training (phish info time) and each email (median time/e-mail). The time spent on the phishing training was very skewed, so a log transformation was used to normalize the data.

As part of their enrollment in the SBO, participants in Experiment 2 agreed to install monitoring software on their home computers that tracked all of their activity, including Internet browsing, installed applications, processes, network connections, system events, and more. These data were used to determine whether participants had been exposed to malicious websites or malicious files and to describe their overall risk of exposure, in terms of the intensity of their browsing and their downloading activity. Malicious websites were identified using the Google Safe Browsing API with participants’ network packet data, which include all HTTP traffic for each webpage. This process captures malicious ads and images embedded within a legitimate webpage separately. Malicious files were identified with VirusTotal.com, a subsidiary of Google that aggregates over 70 anti-virus scanners. A file was considered malicious if more than one scanner flagged it. Using greater scanner agreement did not significantly change the results. We assessed each outcome as a binary variable (where 1 indicates that the outcome was observed at least once and 0 indicates that the outcome was not observed), rather than a continuous one (i.e. number of negative outcomes) due to the unreliability of count data caused by technical bugs in the monitoring software (Canfield et al. 2017).

The primary predictors for the real-world outcomes were (a) intensity of browsing and (b) downloading activity. Browsing intensity combines total URLs visited per day, unique URLs visited per day, and domains visited per day with Cronbach's alpha of 0.79. Given the large skew, a log transformation was used. Downloading activity was measured as a count of the total software installed on the computer. This variable was also highly skewed, so a log transformation was used (Canfield et al. 2017).

Stimuli

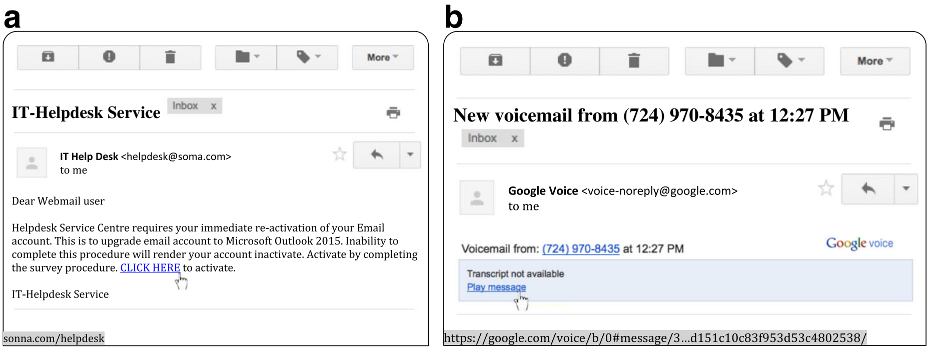
Of the 40 emails that participants reviewed, 19 were phishing emails (adapted from public archives), 19 were legitimate emails (adapted from real ones), and 2 were attention checks. Although a 50% base rate of phishing emails is not realistic (less than 1% of actual emails are phishing), that rate was used to reduce the burden on participants and the time required to collect sufficient data for analysis. The order of the emails was randomized for each participant.

Each phishing email 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 stated sender and subject, (d) requests for urgent action, and (e) grammatical errors or misspellings (Downs et al. 2006). Although the cues were not systematically varied (to reflect their distribution in phishing emails), the URL was the most valid cue for identifying a phishing e-mail (following Downs et al. 2006). 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., misspellings). Figure 1 shows example phishing and legitimate e-mails used in the study.

In the Fig. 1 examples, the phishing email cues include an impersonal greeting ("Dear Webmail user"), suspicious URL ("sonna.com" rather than soma.com), unusual content ("re-activation of your *Email* account"), requests for urgent action ("inability to complete this procedure will render your account *inactivate*"), and grammatical errors (as depicted in italics in the previous two examples).

Analyses

Metacognition was assessed via the Brier score, a measure of the accuracy of probability (confidence) judgments (Brier 1950). It is the average of the squared difference between the judgment or forecast, represented as f, and the observed outcome, represented as o, across all



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

New voicemail from (724) 970-8435 at 12:27 PM

Google Voice <voice-noreply@google.com>
to me

Voicemail from: (724) 970-8435 at 12:27 PM

Transcript not available
Play message

https://google.com/voice/b/0#message/3...d151c10c83f953d53c4802538/

Fig. 1 Example (**a**) phishing and (**b**) legitimate email used in the study

emails for each individual, where smaller values indicate better forecasts. In this context, each individual is forecasting (via their confidence judgment) their accuracy for detecting phishing emails. The Brier score can be decomposed into three components, calibration (or reliability), resolution, and knowledge (or uncertainty). Probability judgments are typically binned into ranges.

$$\text{Brier score} = \frac{1}{N}\sum\nolimits_{t=1}^{N}(f_t - o_t)^2 = \text{Calibration}-\text{Resolution} + \text{Knowledge}$$

$$\text{Calibration} = \frac{1}{N}\sum\nolimits_{k=1}^{K} n_k(f_k - o_k)^2$$

$$\text{Resolution} = \frac{1}{N}\sum\nolimits_{k=1}^{K} n_k(o_k - o\ )^2$$

$$\text{Knowledge} = o(1 - o\ )$$

Following convention, K = 6 bins, defined as 50–59%, 60–69%, 70–79%, 80–89%, 90–99%, and 100%. For each individual, f_k is the average of their confidence ratings within that bin. Then, $\overline{o}_k$ is the outcome mean in each bin k (i.e. the sum of correct forecasts divided by the total forecasts reported for that confidence bin); $\overline{o}$ is the overall outcome mean; and $N = 38$ for the number of forecasts per person. The best score for calibration is 0, when the forecasted frequency and realized frequency are equal. Thus, in the best-case scenario, $f_k = \overline{o}_k$ in each bin, indicating for example, that when the individual is 50% confident, they are correct 50% of the time. In contrast, poor calibration would be high confidence with low or no accuracy and low confidence with high or perfect accuracy. For example, assuming equal sample sizes within each bin, a poor calibration score would be $(.5 - 1)^2 + (.6 - 1)^2 + (.7 - 1)^2 + (.8 - 0)^2 + (.9 - 0)^2 + (1 - 0)^2 = 2.95$. This might happen if participants reported that all emails were phishing emails and assigned all of the phishing emails low confidence. The worst-case calibration score would be if the individual had 0% accuracy across all confidence bins.

The worst score for resolution is 0, when the observed outcomes are the same across all bins so $\overline{o}_k = \overline{o}$. This indicates that individuals have the same accuracy ($\overline{o}_k$) regardless of the confidence bin. In the best-case scenario, individuals would be able to resolve their accuracy into separate confidence bins. Thus, assuming 6 forecasts within each of K confidence bins and an overall accuracy of $(1 + 2 + 3 + 4 + 5 + 6)/(6 * K) = .58$, resolution would be $(.17 - .58)^2 + (.33 - .58)^2 + (.5 - .58)^2 + (.67 - .58)^2 + (.83 - .58)^2 + (1 - .58)^2 = 0.48$.

Knowledge describes the underlying uncertainty of the outcome, where more uncertain events are more difficult to forecast. As Knowledge is not an element of metacognition, we will not consider it further. Brier, calibration, and resolution scores were calculated for all stimuli and separately for phishing and legitimate emails. For comparison, we also include overconfidence defined as the signed difference between mean confidence rating and proportion of correct answers. Positive values indicate over-confidence and negative values underconfidence. To avoid redundancy, overconfidence is omitted from most of the analyses.

Correlations and regression analyses are reported below to explain the relationships between metacognition metrics (calibration and resolution scores), individual and task factors, and real-world vulnerability.

Results

Results are reported below by research question, rather than by experiment, in order to address each question with all available data.

Metacognition for phishing versus legitimate emails

Table 1 reports summary statistics for metacognition metrics by experiment and separately for phishing and legitimate emails. On average, participants had high confidence in their detection judgments in both experiments, consistent with previous research (Fleming and Lau 2014). There was no significant difference between Experiment 1 and 2 for confidence or accuracy, t-tests, $p > 0.05$. Across all questions and individuals, mean confidence was 85% ($SD = 8\%$). Mean accuracy was much lower at 67% ($SD = 11\%$), making for aggregate overconfidence of 18% (=85%–67%). On average, participants were more confident for phishing than legitimate emails, paired t-test, $t(249) = -3.78$, $p < 0.001$.

The metacognitive metrics across all emails were similar to those reported by Li et al. (2016) for calibration ($M = 0.08$ vs. 0.09_{Li}), resolution ($M = 0.03$ vs. 0.04_{Li}), and the Brier scores ($M = 0.26$ vs. 0.25_{Li}). Overall, calibration was generally good, as scores closer to 0 are better, indicating that participants had some knowledge about the accuracy of their judgments. However, resolution was very poor.

Figure 2a shows calibration curves for all responses in the 2 experiments, which were very similar. For the (few) cases where participants expressed low confidence, their judgments were well-calibrated, in the sense of mean confidence being close to proportion correct. As confidence increased, calibration decreased, emerging as overconfidence, such that, for example, participants were correct only 70% of the time when they were 90–99% confident. Participants also had poor resolution, as demonstrated by the flatness of the curve. Particularly when participants were less than 90% confident, they demonstrated little ability to resolve their performance into separate confidence bins and accuracy is relatively constant.

Phishing and legitimate emails reveal different patterns. Calibration was much worse for phishing emails ($M = 0.18$) than for legitimate ones ($M = 0.09$), paired t-test, $t(249) = 6.65$, $p < .001$. Most participants had good calibration (close to 0) for both kinds of email. A few had poor calibration for one kind of email, but not both. Resolution was better for phishing emails ($M = 0.07$) than for legitimate emails ($M = 0.04$), paired t-test, $t(249) = 6.95$, $p < .001$. All

Table 1 Mean and standard deviation (in parentheses) of metacognitive metrics for phishing, legitimate, and all emails

	All emails (38 items)			Legitimate emails (19 items)			Phishing emails (19 items)		
	All	Exp 1	Exp 2	All	Exp 1	Exp 2	All	Exp 1	Exp 2
Confidence	0.85 (0.08)	0.85 (0.08)	0.85 (0.08)	0.85 (0.09)	0.84 (0.08)	0.85 (0.09)	0.86 (0.08)	0.86 (0.08)	0.86 (0.09)
Calibration	0.08 (0.05)	0.07 (0.05)	0.08 (0.06)	0.09 (0.11)	0.08 (0.08)	0.11 (0.14)	0.18 (0.16)	0.19 (0.16)	0.17 (0.16)
Resolution	0.03 (0.02)	0.03 (0.02)	0.03 (0.02)	0.04 (0.04)	0.04 (0.04)	0.04 (0.03)	0.07 (0.05)	0.07 (0.05)	0.06 (0.04)
Brier Score	0.26 (0.08)	0.25 (0.08)	0.26 (0.09)	0.20 (0.14)	0.18 (0.11)	0.22 (0.16)	0.32 (0.17)	0.32 (0.17)	0.31 (0.16)
Accuracy	0.67 (0.11)	0.67 (0.11)	0.67 (0.11)	0.76 (0.19)	0.77 (0.17)	0.74 (0.21)	0.58 (0.20)	0.56 (0.20)	0.59 (0.20)
Cronbach's Alpha	0.59	0.60	0.59	0.77	0.78	0.77	0.77	0.73	0.81
Judges	250	152	98	250	152	98	250	152	98
Forecasts	9,500	5,776	3,724	9,500	5,776	3,724	9,500	5,776	3,724

participants had poor resolution (i.e. close to 0), with no relationship between resolution for phishing and legitimate emails.

Figure 2b displays the calibration curves separately for legitimate and phishing emails, which show quite different patterns. Participants were consistently overconfident for phishing emails, with their judged probability of being correct higher than their actual probability, except for the few times when they said that they were guessing (50%). For example, they correctly identified only 56% of the phishing messages when they were 90–99% confident. Indeed, their proportion correct was barely related to their confidence, except for expressions of certainty (100%). In contrast, participants were relatively well calibrated for legitimate

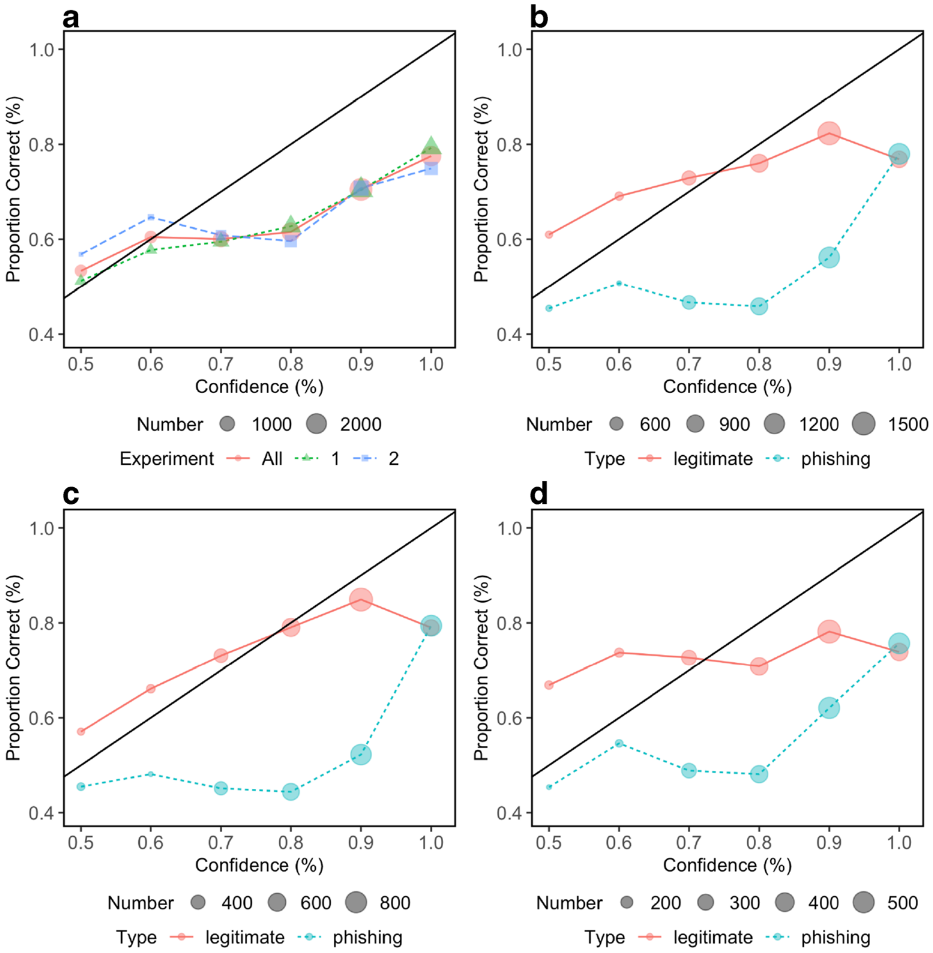


Fig. 2 Calibration curves for (**a**) all emails by experiment, (**b**) legitimate and phishing emails separately across both experiments, (**c**) legitimate and phishing emails separately for Experiment 1, and (**d**) legitimate and phishing emails separately for Experiment 2. Although participants were better calibrated for legitimate emails, they tended to be over-confident for phishing emails. Overall resolution was poor, as demonstrated by the flatness of the curves. The size of the dots indicates the number of observations in that confidence bin

emails, except when they expressed 100% confidence. Participants had better resolution for phishing emails, primarily due to the larger difference in accuracy between low confidence (50%) and high confidence (100%) judgements. Thus, although participants were generally overconfident for phishing emails, they did demonstrate an ability to distinguish between cases when they should have low or high confidence. The pattern of results suggests that anything less than 90% confident could be considered "low confidence."

Cronbach's alpha for detection accuracy was higher for legitimate ($\alpha = 0.77$) and phishing ($\alpha = 0.77$) emails considered separately, than for all the emails ($\alpha = 0.59$), indicating greater consistency within the two categories than overall. Kleitman et al. (2018), observed similar internal consistency, $\alpha = 0.81$ for legitimate email detection and $\alpha = 0.80$ for phishing email detection (evaluating all emails was outside the scope of their analysis). These differences support analyzing phishing and legitimate emails separately. Across all emails, there was little difference between the two experiments reported here. Overall, legitimate, and phishing calibration were not significantly different between Experiments 1 and 2, $p > .05$. This suggests it is appropriate to combine the experimental results in this analysis.

Figure 3a displays each email in terms of its actual difficulty (proportion correct) and perceived difficulty (mean confidence). There was much greater variance in the difficulty of the phishing emails than the legitimate ones. Almost all of the legitimate emails were correctly identified as such by 60%–90% of participants. However, for the phishing emails, the proportion correctly identified ranged from 15% to 85%. Thus, some were much more deceptive than others. There was little variance in participants' mean confidence when evaluating individual emails, in either category (ranging from 80 to 90% confident). Thus, some of the phishing emails not only fooled participants, but also left them unaware that they were being fooled. In contrast, legitimate emails evoked suspicions that were roughly proportionate to their chance of being warranted. As shown in Table 1, participants were more miscalibrated for phishing emails than for legitimate ones. This pattern was found in both experiments (see Fig. 3b, c).

Individual and task factors influencing metacognition

Figure 4 displays each participant in terms of mean confidence and proportion of correct responses. Although they varied widely in confidence and accuracy, almost all were overconfident, as seen in their falling below the diagonal. There were no discernible differences between the two experiments.

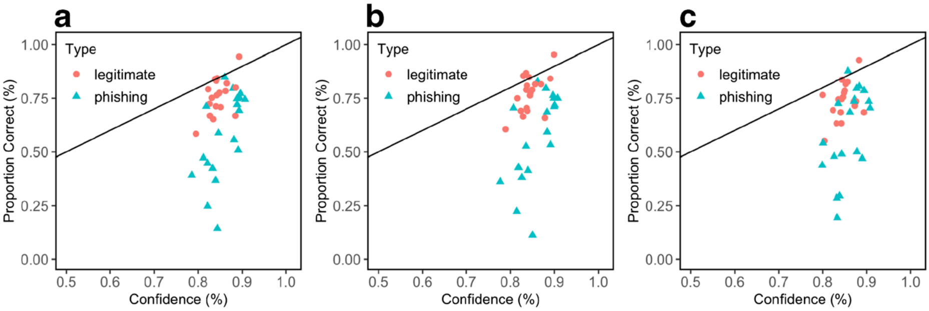


Fig. 3 Observations averaged across legitimate and phishing emails separately across (**a**) all experiments, (**b**) Experiment 1, and (**c**) Experiment 2

Table 2 summarizes linear regressions predicting calibration and resolution, for all emails, legitimate emails, and phishing emails. The strongest effect was for perceived consequences on the calibration for legitimate and phishing emails. For legitimate emails, participants who thought that falling for phishing attacks had greater (negative) consequences were less well calibrated (i.e. calibration was further from 0). In contrast, for phishing attacks, they were better calibrated. None of the predictors were significantly related to resolution ($p > .01$).

Relationship between metacognition and real-world vulnerability

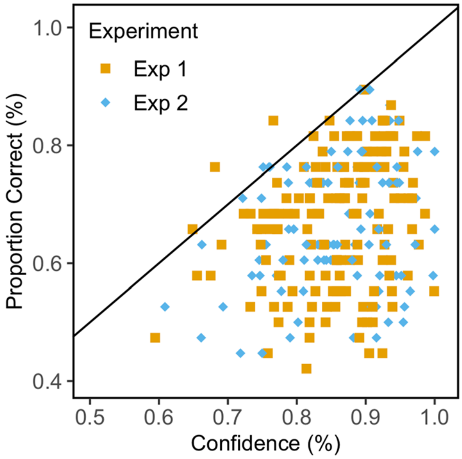
In Experiment 2, performance in the online phishing detection experiment can be compared to real-world vulnerability. Most users (84/93 = 90%) had malicious files on their computer, which was more likely for those with poorer resolution, both overall, $r(93) = -0.27$, $p = .009$, and for phishing emails specifically, $r(93) = -0.21$, $p = .04$. None of the metacognitive measures were correlated to participants' browsing intensity or likelihood of having malicious URLs on their home computers, $p > .05$. Resolution of all emails was more correlated with resolution for phishing emails, $r(98) = 0.56$, $p < .001$, than legitimate emails, $r(98) = 0.23$, $p = .02$. This may have been because resolution was slightly higher, and thus more variable, for phishing emails than legitimate emails.

Table 3 reports logistic regression analyses predicting the presence of malicious files from the metacognitive measures in Experiment 2. Model 1 evaluated the effect of adding calibration to logistic regressions for all, legitimate, and phishing emails. The LRTs were not significant, indicating that calibration does not improve model fit. However, consistent with Canfield et al. (2017), the total number of software downloads was a significant predictor of the odds of malicious files, for all sets of messages.

Model 2 evaluated the effect of adding resolution to the three logistic regressions. For all emails, overall resolution was a weakly significant predictor of the odds of malicious files. As resolution improved (i.e. increased), the odds of malicious files were lower. When resolution was included in the model, age was also a significant predictor, with younger participants being more likely to have malicious files on their home computers.

Model 3 evaluated the effect of adding both metacognition metrics, resolution and calibration, to the logistic regressions. As with Model 2, resolution and age were significant

Fig. 4 Observations averaged across each individual, separated by Experiment



predictors for all emails. For legitimate emails, both calibration and resolution were significant predictors. This is inconsistent with Model 1, where calibration did not improve model fit. Participants with better legitimate email calibration had worse legitimate email resolution (i.e. close to 0), $r(98) = 0.26$, $p = .009$. In general, Models 2 and 3 for all emails had the best model fit and lowest AIC values. The inconsistent results for calibration between Models 1 and 3 suggest that it may not be a true predictor.

Discussion

Understanding the relationship between metacognition and phishing detection is critical for improving training and education. These analyses sought to identify (1) how metacognition differs for phishing and legitimate emails, (2) the relationship between metacognition and individual/task factors for phishing and legitimate emails, and (3) the relationship between metacognition and real-world vulnerability. Few studies have examined metacognition in the context of phishing detection, and none have examined it separately for phishing and legitimate emails. This is also the first study to examine the relationship between metacognition and real-world computer security vulnerability.

Overall, the research has three main findings, described in greater depth below. First, we found that metacognitive performance differed for phishing and legitimate emails. Cronbach's alpha was higher for legitimate and phishing emails considered separately, rather than for all emails, indicating that different processes were involved. Overall, participants had relatively good calibration for legitimate emails (Fig. 2b, c). Thus, participants generally assigned the

Table 2 Linear regression analysis for calibration and resolution, combining Experiments 1 and 2.

	Calibration			Resolution		
	All	Legitimate	Phishing	All	Legitimate	Phishing
	B	B	B	B	B	B
	(SE)	(SE)	(SE)	(SE)	(SE)	(SE)
Intercept	0.07	−0.16*	0.54***	0.05**	0	0.09*
	(0.04)	(0.08)	(0.12)	(0.02)	(0)	(0.04)
Passed attention check	−0.03*	−0.05*	−0.04	0	0	0.01
	(0.01)	(0.02)	(0.03)	(0)	(0)	(0.01)
log(Phish info time)	0	−0.01	0.01	0	0	0.01*
	(0)	(0.01)	(0.01)	(0)	(0)	(0)
Median time/email	−0.01	−0.02	−0.02	0	0	0
	(0.01)	(0.01)	(0.02)	(0)	(0)	(0.01)
Average perceived consequences	−0.01	0.05***	−0.07***	0	0	0
	(0.01)	(0.01)	(0.01)	(0)	(0)	(0)
log(Age)	0.02	0.05*	−0.02	0	0	−0.01
	(0.01)	(0.02)	(0.03)	(0)	(0)	(0.01)
Male	−0.01	0	−0.02	0	0	0
	(0.01)	(0.01)	(0.02)	(0)	(0)	(0.01)
College	−0.01	−0.03*	0.01	0	0	0
	(0.01)	(0.01)	(0.02)	(0)	(0)	(0.01)
N	231	231	231	231	231	231
Adjusted R^2	0.03	0.16	0.08	0	−0.01	0
F-test	*F*(7,223)=	*F*(7,223)=	*F*(7,223)=	*F*(7,223)=	*F*(7,223)=	*F*(7,223)=
	2.03	7.30***	3.85***	0.94	0.80	1.12

The asterisks indicate statistical significance, where * is $p < .05$, ** is $p < .01$, and *** is $p < .001$

Table 3 Logistic regression analysis for the presence of malicious files for Experiment 2.

	Model 1 – Calibration			Model 2 – Resolution			Model 3 – Metacognition		
	B (SE)			B (SE)			B (SE)		
	All	Legitimate	Phishing	All	Legitimate	Phishing	All	Legitimate	Phishing
Intercept	−7.38	−6.33	−7.00	−5.50	−6.80	−5.76	−5.57	−7.19	−6.35
	(3.87)	(3.59)	(3.89)	(4.48)	(3.76)	(3.88)	(4.60)	(4.14)	(4.15)
Calibration	12.66	22.34	1.22				11.56	60.12*	1.60
	(11.25)	(16.58)	(3.37)				(13.67)	(30.53)	(3.42)
Resolution				−82.01*	−11.16	−22.75	−80.81*	−63.99*	−23.45
				(32.50)	(14.69)	(13.01)	(33.41)	(31.25)	(13.27)
log(Total Software)	2.47**	2.00**	2.41**	3.07**	2.52**	2.68**	3.03**	2.72*	2.81**
	(0.85)	(0.73)	(0.86)	(1.16)	(0.91)	(0.97)	(1.13)	(1.06)	(1.03)
Age	−0.05	−0.04	−0.04	−0.09*	−0.05	−0.06	−0.10*	−0.07	−0.07
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)
Male	−0.62	−0.88	−0.64	−1.63	−0.69	−1.17	−1.55	−1.49	−1.05
	(0.91)	(0.95)	(0.91)	(1.14)	(0.89)	(0.96)	(1.20)	(1.21)	(0.99)
College	−1.34	−0.93	−1.03	0.17	−0.90	−0.65	−0.32	−0.73	−0.76
	(1.27)	(1.22)	(1.22)	(1.37)	(1.19)	(1.23)	(1.54)	(1.29)	(1.25)
N	91	91	91	91	91	91	91	91	91
AIC	46.46	44.39	47.80	38.33	47.37	44.47	39.50	39.81	46.23
LRT	$\chi^2(1)$ = 1.48	$\chi^2(1)$ = 3.55	$\chi^2(1)$ = 0.14	$\chi^2(1)$ = 9.61**	$\chi^2(1)$ = 0.57	$\chi^2(1)$ = 3.47	$\chi^2(2)$ = 10.44**	$\chi^2(2)$ = 10.13**	$\chi^2(2)$ = 3.71

The Likelihood Ratio Test (LRT) compares the full models with models excluding calibration and resolution. The asterisks indicate statistical significance, where * is $p < .05$, ** is $p < .01$, and *** is $p < .001$

"right" confidence to their detection decisions. However, participants were generally overconfident and exhibited a metacognitive (high confidence) bias for phishing emails. The detection task was significantly more difficult for phishing emails, as evidenced by the larger range of observed accuracies (Fig. 3). Given that difference in difficulty, the calibration curves look much like those for easy and hard tasks, as first observed by Lichtenstein and Fischhoff (1977). Participants had very poor resolution, although it was slightly better for phishing than legitimate emails, indicating that they struggled to distinguish between judgments that were correct and incorrect. Yates (1982) argues that resolution is far more important than calibration for improving forecasts, as it reflects a skill that is informative, rather than probabilistic, at the levels of individual judgments.

Second, we found large variation in individual metacognitive performance. The individual and task measures used in this research were not useful predictors (Table 2). Attention to the task, time spent reviewing anti-phishing training information, time spent per email, and demographic variables were all not significant predictors of calibration or resolution ($p > .01$). This was inconsistent with earlier research, which found that participants who spent more time per email have better calibration (Li et al. 2016). Our only significant finding was that participants who thought that falling for phishing attacks had greater (negative) consequences were better calibrated for phishing emails and less well calibrated for legitimate emails. This was likely due to a bias toward perceiving emails as phishing, which improved accuracy for phishing emails, while reducing accuracy for legitimate emails.

Lastly, the SBO provided a unique opportunity to compare performance on an experimental task with actual experience. We found no relationship between metacognitive performance and browsing intensity or likelihood of visiting a malicious website. However, we found a weak, but statistically significant tendency for participants with better resolution to be less likely to have malicious files on their home computers ($p < .05$), even after controlling for risk of exposure to cyber threats via overall download activity. Although, this test was limited by the fact that so few participants had no malicious files (9 of 98), providing limited variance to predict.

To the extent that these results generalize, we believe that they reflect the feedback that people receive on their email judgments and its effects on their ability to learn how to detect phishing emails. With legitimate emails, people receive feedback from replies and interactions in the real world that confirm the emails' legitimacy. With phishing emails, people may not realize that they have acquired a virus or had their identity stolen until long after a misidentified message; even then, it may be difficult to know which message was the source. If the individual was not the direct target, but rather the portal through which an intruder gained entry to a system and accomplished their mischief elsewhere, that individual may never learn about their mistake. Even if someone receives a notification from their IT department indicating that a particular email was malicious, they may have difficulty remembering what they were thinking when they originally clicked on it. The fact that participants were more successful at detecting phishing emails when they were 100% confident suggests that they have learned heuristics that apply to that case (e.g., their bank does not ask for account information via email). Conceivably, that knowledge may convey an unwarranted feeling of general sophistication when dealing with other messages.

Limitations

There are several limitations to this work. First, our stimuli used a 50% base rate for phishing emails, which is much higher than in real life (typically <1%). For such tasks, research suggests that performance improves as the base rate increases (Wolfe et al. 2007). We are also drawing participants' attention to phishing. As a result, the performance reported here represents an upper limit of what would be expected in the field.

In addition, the real-world vulnerability measures were very noisy. Exposure to malicious events can be affected by how new a computer is, which browsers users chose to use (perhaps for non-security reasons), and whether they have enabled automatic updates (Canfield et al. 2017). As a result, any effect may be small. Moreover, our sample size was limited by the number of current SBO participants. Future work might revisit the SBO, which now has a larger sample and improved data collection software, perhaps making it possible to use count data, rather than binary outcome measures, which would also improve statistical power.

Implications for interventions

Computer users receive poor feedback on how well they are doing and how much confidence to place in their judgments. Research on vigilance interventions has found that even artificial injections of feedback can improve performance on real-world tasks. For example, "signal injection and performance feedback" has improved vigilance in sonar watchstanding (Mackie et al. 1994), baggage security screening (Wolfe et al. 2013, 2007), and medical diagnosis (Evans et al. 2013). It entails artificially injecting tests throughout normal performance of a task and giving feedback on whether participants' success. Wolfe et al. (2007, 2013) found that exposing baggage screeners to brief bursts of such training improved their detection ability, even after they returned to a real-world setting with a low base rate of signals and no feedback. In addition to providing feedback, such tests serves to increase the base rate of signals, leading people to perceive more stimuli as signals (Goodie and Fantino 1999; Kluger and DeNisi 1996). It is possible that such an intervention may also be valuable for improving metacognition. More research is needed to measure an effect.

In the context of education and self-regulated learning, error management has emerged as an effective method, with parallels to feedback-based training. It emphasizes learning from errors, rather than aiming to avoid errors, by drawing on emotional self-regulation and metacognition. Essentially, allowing people to learn from experience through errors, rather than from instruction, encourages metacognitive thinking and emotional control. It has been found to improve performance in a subsequent task that requires new solutions (Keith and Frese 2005). Some anti-phishing training has a similar philosophy. A popular workplace intervention is to send artificial phishing emails to employees, called "embedded training" (e.g. Wombat Security, PhishMe). If employees fall for those emails, they are told so, resulting in increased performance on detecting subsequent phishing emails (Kumaraguru et al. 2010). However, it provides only partial feedback. Individuals who never click on the artificial phishing emails never receive notice that they made the right choice. It may be valuable to test the effect of providing full feedback in order to support self-regulated learning. Mohan et al. (2017, 2018) report success with a similarly conceived intervention aimed at emergency department physicians' triage decisions.

Phishing detection is a subset of digital literacy that K-12 educators could include as part of curricula on critical thinking and information literacy. For example, they might use games like Anti-Phishing Phil for learning principles of scientific hypothesis testing in the context of learning how to detect fraudulent websites (Sheng et al. 2007). Information literacy may also be considered a basic skill to be taught on its own (Johnston and Webber 2003). Colleges and universities are already implementing embedded training to reduce phishing vulnerability. It may be valuable to leverage insights from self-regulated learning to better design these trainings to improve metacognitive abilities.

Given the disparity observed here, in metacognitive performance on legitimate and phishing emails, we believe that a promising path for future research is to investigate which features of phishing messages make them so deceptive. It may be valuable to perform within-subject analysis of the impact of specific features on performance and metacognition. This may aid efforts to help users understand which kinds of phishing emails they are more vulnerable to, without making them overconfident in their ability to detect those messages. Future studies developing anti-phishing interventions should investigate both performance and metacognition metrics to understand how users are shifting their decision-making strategies. If successful, an intervention might make the calibration curve for phishing messages look more like the curve for legitimate emails. In addition, if attackers changed their strategies, the calibration curve for phishing messages might retain its current shape, but have a much higher share of messages with low confidence (50%).

Acknowledgements The Security Behavior Observatory was partially funded by the NSA Science of Security Lablet at Carnegie Mellon University (contract #H9823014C0140); the National Science Foundation, Grant CNS-1012763 (Nudging Users Towards Privacy); and the Hewlett Foundation, through the Center for Long-Term Cybersecurity (CLTC) at the University of California, Berkeley. In addition, this work was supported by the Swedish Foundation for the Humanities and Social Sciences and Riksbankens Jubileumsfond. We thank three anonymous reviewers for their thoughtful comments.

Compliance with ethical standards The authors have no conflicts of interest. This research involved human participants and was approved by the Carnegie Mellon University Institutional Review Board. All participants provided informed consent.

References

Blackshaw, L., & Fischhoff, B. (1988). Decision making in online searching. *Journal of the American Society for Information Science, 39*(6), 369–389.

Bond, C. F., & Depaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*(3), 214–234 Retrieved from https://www.aclu.org/sites/default/files/field_document/2006-Personality-and-Social-Psychology-Review-Accuracy-of-Deception-Judgements.pdf.

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and Bias. *Psychological Bulletin, 134*(4), 477–492. https://doi.org/10.1037/0033-2909.134.4.477.supp.

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 the Human Factors and Ergonomics Society*, 1115–1119. https://doi.org/10.1177/1071181311551233.

Brier, G. W. (1950). Verification of forecasts expressing probability. *Monthly Weather Review, 78*, 1–3.

Canfield, C., Fischhoff, B., & Davis, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. *Human Factors, 58*(8), 1158–1172. https://doi.org/10.1177/0018720816665025.

Canfield, C., Davis, A., Fischhoff, B., Forget, A., Pearman, S., & Thomas, J. (2017). Replication: Challenges in using data logs to validate phishing detection ability metrics. In *Symposium on Usable Privacy and Security* (pp. 271–284). Retrieved from https://www.usenix.org/conference/soups2017/technical-sessions/presentation/canfield

Cranor, L. F. (2008). A framework for reasoning about the human in the loop. *Proceedings of the 1st Conference on Usability, Psychology, and Security*, 1:1–1:15. https://doi.org/10.1109/MSP.2010.198.

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The accuracy-confidence correlation in the detection of deception. *Personality and Social Psychology Review*, *1*(4), 346–357. Retrieved from http://www.ffri.hr/~ibrdar/komunikacija/seminari/DePaulo, 1997 - Detection of deceiption . meta-analysis.pdf.

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science, 29*(5), 761–778. https://doi.org/10.1177/0956797617744771.

Dinsmore, D. L., Alexander, P., & Loughlin, S. M. (2008). Focusing the conceptual Lens on metacognition, self-regulation, and self-regulated learning learning. *Educational Psychology Review, 20*, 391–409. https://doi.org/10.1007/s10648-008-9083-6.

Downs, J. S., Holbrook, M. B., & Cranor, L. F. (2006). Decision strategies and susceptibility to phishing. *Proceedings of the Second Symposium on Usable Privacy and Security - SOUPS '06*, *15213*, 79. https://doi.org/10.1145/1143120.1143131.

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system? *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, 2399. https://doi.org/10.1145/1753326.1753688.

Eshet-Alkalai, Y. (2004). Digital literacy: A conceptual framework for survival skills in the digital era. *Journal of Educational Multimedia and Hypermedia, 13*(1), 93–106.

Evans, K. K., Birdwell, R. L., & Wolfe, J. M. (2013). If you Don't find it often, you often Don't find it: Why some cancers are missed in breast Cancer screening. *PLoS One, 8*(5), 1–6. https://doi.org/10.1371/journal.pone.0064366.

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

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 1–9. https://doi.org/10.3389/fnhum.2014.00443.

Forget, A., Komanduri, S., Acquisti, A., Christin, N., Cranor, L. F., & Telang, R. (2014). Security behavior observatory : Infrastructure for long- term monitoring of client machines security behavior observatory : Infrastructure for long-term monitoring of client machines.

Forget, A., Pearman, S., Thomas, J., Acquisti, A., Christin, N., Cranor, L. F., … Telang, R. (2016). Do or do not, there is no try: User engagement may not improve security outcomes. *Proceedings of the Twelfth Symposium on Usable Privacy and Security (SOUPS)*, (Soups), 97–111. Retrieved from https://www.usenix.org/conference/soups2016/technical-sessions/presentation/forget

Goodie, A. S., & Fantino, E. (1999). What does and does not alleviate base-rate neglect under direct experience. *Journal of Behavioral Decision Making, 12*(4), 307–335. https://doi.org/10.1002/(SICI)1099-0771(199912)12:4<307::AID-BDM324>3.0.CO;2-H.

Greene, J. A., Yu, S. B., & Copeland, D. Z. (2014). Measuring critical components of digital literacy and their relationships with learning. *Computers & Education, 76*, 55–69. https://doi.org/10.1016/j.compedu.2014.03.008.

Hauch, V., Sporer, S. L., Michael, S., & Meissner, C. A. (2016). Does training improve the detection of deception? *Communication Research, 43*(3), 283–343. https://doi.org/10.1177/0093650214534974.

Hodgin, E., & Kahne, J. (2018). Misinformation in the information age: What teachers can do to support students. *Social Education, 82*(4), 208–211 Retrieved from http://eddaoakland.org/wp-content/.

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM, 50*(10), 94–100. https://doi.org/10.1145/1290958.1290968.

Johnston, B., & Webber, S. (2003). Information literacy in higher education: A review and case study. *Studies in Higher Education, 28*(3), 335–352. https://doi.org/10.1080/03075070309295.

Keith, N., & Frese, M. (2005). Self-regulation in error management training: Emotion control and metacognition as mediators of performance effects. *Journal of Applied Psychology, 90*(4), 677–691. https://doi.org/10.1037/0021-9010.90.4.677.

Kleitman, S., Law, M. K. H., & Kay, J. (2018). It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling. *PLoS One, 13*(10), 1–29.

Kluger, A. N., & DeNisi, A. (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037//0033-2909.119.2.254.

Koltay, T. (2011). The media and the literacies: Media literacy, information literacy, digital literacy. *Media, Culture and Society, 33*(2), 211–221. https://doi.org/10.1177/0163443710393382.

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. https://doi.org/10.1145/1754393.1754396.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition, 10*, 294–340. https://doi.org/10.1006/ccog.2000.0494.

Law, M. K. H., Jackson, S. A., Aidman, E., Geiger, M., Olderbak, S., & Kleitman, S. (2018). It's the deceiver, not the receiver: No individual differences when detecting deception in a foreign and a native language. *PLoS One, 13*(5), 1–17. https://doi.org/10.1371/journal.pone.0196384.

Lee, N. M. (2018). Fake news, phishing, and fraud: A call for research on digital media literacy education beyond the classroom. *Communication Education, 67*(4), 460–466 Retrieved from https://illiad.mst.edu/illiad/illiad.dll?Action=10&Form=75&Value=238353.

Li, Y., Wang, J., & Rao, H. R. (2016). An examination of the calibration and resolution skills in phishing email detection. Americas conference on information systems. Retrieved from http://repository.ittelkom-pwt.ac.id/1339/1/An examination of the calibration and resolution skills in phishi.Pdf.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human, 183*(3052), 159–183. https://doi.org/10.1016/0030-5073(77)90001-0.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26*(2), 149–171. https://doi.org/10.1016/0030-5073(80)90052-5.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: State of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press.

Macgregor, D., Fischhoff, B., & Blackshaw, L. (1987). Search success and expectations with a computer Interface. *Information Processing & Management, 23*(5), 419–432 Retrieved from http://www.gwern.net/docs/statistics/decision/1987-macgregor.pdf.

Mackie, R. R., Wylie, C. D., & Smith, M. J. (1994). Countering loss of vigilance in sonar watchstanding using signal injection and performance feedback. *Ergonomics, 37*(7), 1157–1184. https://doi.org/10.1080/00140139408964895.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015). Identifying and cultivating Superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science, 10*(3), 267–281. https://doi.org/10.1177/1745691615577794.

Mohan, D., Farris, C., Fischhoff, B., Rosengart, M.R., Angus, D., Yealy, D., Wallace, D., & Barnato, A. (2017). Testing the efficacy of a video game vs. a traditional education program at improving physician decision making in trauma triage: A randomized controlled trial. *BMJ, 359,* j5416. MJ2017;359:j5416.

Mohan, D., Fischhoff, B., Angus, D. C., Rosengart, M. R., Wallace, D. J., Yealy, D. M., Farris, C., Chang, C.-C. H., Kerti, S., & Barnato, A. E. (2018). Serious video games may improve physicians' heuristics in trauma triage. *PNAS, 115*(37), 9204–9209. https://doi.org/10.1073/pnas.1805450115.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making, 5*(5), 411–419. https://doi.org/10.2139/ssrn.1626226.

Pattinson, M., Jerram, C., Parsons, K., McCormac, A., & Butavicius, M. (2012). Why do some people manage phishing e-mails better than others? *Information Management and Computer Security, 20*(1), 18–28. https://doi.org/10.1108/09685221211219173.

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. https://doi.org/10.1177/0018720815585906.

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 *Symposium on Usable Privacy and Security* (pp. 88–99). Retrieved from http://cups.cs.cmu.edu/antiphishing_phil/

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L. F., & Downs, J. (2010). Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. Proceedings of the 28th international conference on human factors in computing systems - CHI '10, 373–382. https://doi.org/10.1145/1753326.1753383.

Smith, D. J., Shields, W. E., & Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences, 26*, 317–373 Retrieved from http://psychweb.psy.umt.edu/faculty/shields/shields.html.

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*(1), 3–14. https://doi.org/10.1007/s11409-006-6893-0.

Verizon. (2018). *2018 Data Breach Investigations Report*. Retrieved from https://enterprise.verizon.com/resources/reports/dbir/

Von Hippel, W., Baker, E., Wilson, R., Brin, L., & Page, L. (2016). Detecting deceptive behaviour after the fact. *British Journal of Social Psychology, 55*, 195–205. https://doi.org/10.1111/bjso.12129.

Vrij, A., Anders Granhag, P., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest, 11*(3), 89–121. https://doi.org/10.1177/1529100610390861.

Wang, J., Li, Y., & Rao, H. R. (2016). Overconfidence in phishing email detection. *Journal of the Association for Information Systems, 17*(11), 759–783.

Werlinger, R., Hawkey, K., & Beznosov, K. (2009). An integrated view of human, organizational, and technological challenges of IT security management. *Information Management and Computer Security, 17*(1), 4–19. https://doi.org/10.1108/09685220910944722.

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

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), 33. https://doi.org/10.1167/13.3.33.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance, 30*, 132–156 Retrieved from https://deepblue.lib.umich.edu/bitstream/handle/2027.42/23907/0000150.pdf?sequence=1&isAllowed=y.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B, 367*, 1310–1321. https://doi.org/10.1098/rstb.2011.0416.