COMPUTERS & SECURITY 77 (2018) 128–137

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Evaluating the applicability of the double system lens model to the analysis of phishing email judgments

Kylie A. Molinaro[a,b], Matthew L. Bolton[a,*]

[a] Department of Industrial and Systems Engineering, University at Buffalo, State University of New York, Buffalo, NY, USA

[b] Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA

ARTICLE INFO

Article history:
Received 2 December 2017
Revised 21 February 2018
Accepted 30 March 2018
Available online 5 April 2018

Keywords:
Judgment analysis
Lens model
Linear regression
Phishing
Cybersecurity

ABSTRACT

Phishing emails pose a serious threat to cybersecurity. Because human users are the last line of defense, understanding how users identify phishing emails is imperative to addressing this problem. Judgment analysis (JA) provides a means of analyzing both how information in the environment (cues) contributes to an outcome and how users synthesize cues into judgments about that outcome, typically using multiple linear regression. Because JA has not been applied to this domain, this effort assessed if the statistical assumptions of JA with multiple linear regression are upheld. We hypothesized that phishing cues are linearly combinable, meaning a lens model analysis, a type of JA, is appropriate for evaluating phishing judgments. To test this, we analyzed ten participants who judged whether or not emails were phishing using the double system lens model. Results indicated that the lens model is an appropriate means of analyzing phishing judgments, primarily evidenced by the goodness of fits for both the environment model and human judgment models. We also observed varying achievement scores across participants consistent with their varying levels of performance in the judgment task. We discuss our results and how future phishing judgment research can utilize JA afforded analysis capabilities.

1. Introduction

Phishing emails, messages designed to appear legitimate in an attempt to get individuals to reveal personal information or download malicious files, are a serious threat to cybersecurity. Phishing emails generally work by sending individuals a message with a compromised attachment or link, or include wire transfer instructions (Vishwanath et al., 2016). Successful phishing campaigns are an expensive problem, with an estimated annual impact of approximately 2.4 billion dollars (Microsoft, 2014). These expenses are associated with the theft of money, costs associated with identifying and repairing breaches, and the loss of future business. Not only are the numbers of cyber attacks increasing (Passeri, 2016; Volz, 2016), but some of the most damaging data breaches and wire transfer frauds in recent years, like those against Ubiquiti Networks Inc. and the Scoular Co. (Krebs, 2016), began with a phishing attack. The phishing problem continues to grow, with the Anti-Phishing Working Group identifying over 1.2 million separate phishing attacks in 2016, a 65% increase from 2015 (Anti-Phishing Working Group, 2017). Further, Verizon (2017) noted in their 2017 report that 95% of phishing attacks that led to a

* Corresponding author at: Department of Industrial and Systems Engineering, University at Buffalo, State University of New York, Buffalo, NY, USA.
E-mail addresses: kyliemol@buffalo.edu (K.A. Molinaro), mbolton@buffalo.edu (M.L. Bolton).
https://doi.org/10.1016/j.cose.2018.03.012
0167-4048/

breach were followed by software installation, making email attachments the most used delivery vehicle for malware.

Human users will always be the last line of defense against successful email phishing campaigns. Because of this, security groups within organizations often distribute information about how to detect phishing emails to employees. Phishing training and security notices generally focus on describing different phishing cues and where to find them in the email. However, these are not completely effective and even individuals who are informed about basic techniques for recognizing phishing emails can fall for deceptions (Caputo et al., 2014; Davinson and Sillence, 2010; Ferguson, 2005; Hong, 2012; Kumaraguru et al., 2007, 2008).

Clearly, there is a real and urgent need to understand what information humans use when making judgments about whether or not to trust an email so that phishing emails can be appropriately combated. Despite this, very little work has focused on modeling these human judgments (Pfleeger and Caputo, 2012). The work that has been done on this subject has focused on assessing susceptibility based on general individual differences (Williams et al., 2017), individual differences in cognition (Canfield et al., 2016; Vishwanath et al., 2016; Wang et al., 2012), and detection strategies (Downs et al., 2006; Zielinska et al., 2015). However, none of these analyses have focused on understanding how people use information in an email to make judgments about whether or not it constitutes a phishing attempt. The lens model is a statistical modeling judgment analysis technique that allows analysts to understand and predict how people synthesize information sources (cues) into judgments (Brunswik, 1955; Cooksey, 1996). There are a number of known cues that can help indicate if an email is a phishing attempt (Karakasiliotis et al., 2006) This suggests that the lens model would be appropriate for analyzing phishing judgments. However, it has never been used for this purpose.

The majority of lens model analyses rely on multiple linear regression (Karelaia and Hogarth, 2008; Kaufmann et al., 2013). Thus, lens model analyses work well in situations where the information provided by different cues can be linearly combined to make accurate predictions about the criteria on which judgments are being made. In this research, we attempted to evaluate whether or not the multiple linear regression assumptions of the lens model were appropriate for application to the phishing problem.

2. Background

Below we discuss the necessary background for understanding our research on the use of judgment analysis with the lens model in the phishing domain. This includes a survey of the other models that have been used to evaluate human phishing judgments, judgment analysis with the lens model, and information about the cues that appear to be important in phishing judgments.

2.1. *Human models of phishing judgment*

There is deep literature on phishing detection and filtering, however little research has focused on modeling the human user (Pfleeger and Caputo, 2012).

The suspicion, cognition, and automaticity model of phishing susceptibility (SCAM) is a cognitive-behavioral model that aims to measure individual victimization of phishing emails (Vishwanath et al., 2016). The SCAM provides a means of estimating phishing susceptibility based on several factors shown to influence overall suspicion: cyber risk-beliefs, deficient self-regulation, heuristic processing, systematic processing, and email habits. The SCAM questionnaire was administered to participants a week after the phishing email was sent. If the participant recalled the email, they answered Likert scale questions covering all previously listed factors, including overall suspicion.

Using signal detection theory to measure phishing attack vulnerability, Canfield et al. (2016) noted a greater sensitivity was positively correlated with confidence. Greater willingness to treat emails as legitimate was negatively correlated with their actions' perceived consequences and positively correlated with confidence.

Wang et al. (2012) found attention to visceral triggers, attention to phishing deception indicators, and phishing knowledge influenced phishing detection. Cognitive effort did not significantly affect detection likelihood.

Arachchilage and Love (2014) developed a theoretical model to understand how conceptual and procedural knowledge influence a user's self-efficacy against phishing attacks. Their results showed the interaction effect of conceptual and procedural knowledge positively impacted users' self-efficacy, which then enhanced their phishing threat avoidance behavior.

Other researchers have utilized a mental modeling approach. Downs et al. (2006) identified three main strategies participants used when describing their responses to emails: "(1) *this email appears to be for me*, (2) *it's normal to hear from companies you do business with*, (3) *reputable companies will send emails*." The authors noted that the awareness of phishing risks was not linked to perceived vulnerability or to useful strategies, making people more susceptible to phishing attacks. Zielinska et al. (2015) compared the mental model networks of expert and novice computer users. Results indicated experts had more links connecting phishing concepts (such as strategies for preventing phishing, trends and characteristics of phishing attacks, and the consequences of phishing) than novices.

These models help us understand different pieces of the phishing problem, but do not evaluate how a person synthesizes information in their judgments. For this, judgment analysis methods should be appropriate.

2.2. *Judgment analysis*

Judgment analysis (JA), which is based on Brunswick's probabilistic functionalism (Brunswik, 1955), is a technique for analyzing how people make judgments of distal criteria (the environment) using proximal cues (information in the environment) (Cooksey, 1996). While different statistical learning techniques can be used for this purpose (Bruins and Cooksey, 2000; Yoon et al., 2017), the vast majority of lens model analyses are based on multiple linear regression (Karelaia and Hogarth, 2008; Kaufmann et al., 2013). While there are multiple

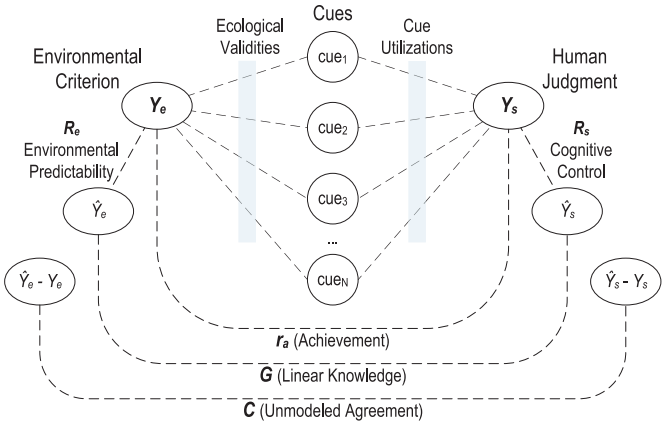

Fig. 1 – Graphical representation of the double system lens model.

versions of JA, this work focuses on the popular double system lens model (Cooksey, 1996).

The double system lens model (Fig. 1) uses symmetric statistical models of the environment and the judgment values made by the human to evaluate human judgment performance. Specifically, the same measurable environmental cues are used as predictors (independent variables) in two fitted linear regression models: one to the criterion (the actual value of the environmental quality that is being judged, $\hat{Y}_e$) and one to judgment values ($\hat{Y}_s$). The weights assigned to the cues (independent variables) allow analysts to compare how differently the cues factor into the prediction of the criterion (ecological validities) and the judgment (cue utilizations). The cue utilizations can be compared between participants to determine how the judgment strategies of each differ. In cases where the cue levels have been normalized, cue weights can also be used to compare the relative weight each cue has in influencing a predicted dependent measure (criterion or judgment).

Further, the lens model equation, originally proposed by Hursch et al. (1964) and later modified by Tucker (1964),

$$r_a = GR_eR_s + C\sqrt{1-R_e^2}\sqrt{1-R_s^2} \quad (1)$$

gives analysts a means of evaluating the achievement of the judge (how well the judge performed on the judgment task) while accounting for the different factors that affect it. r_a is the achievement of the judge represented as the correlation between the criterion (Y_e) and judgment (Y_s). Thus, achievement is measured from low to high by a value between 0 and 1. G represents linear knowledge: a measure of the correspondence between the environment and human judgment model predictions. This is measured as the correlation between $\hat{Y}_e$ and $\hat{Y}_s$. R_e is a measure of the environmental predictability, how well the model of the environment corresponds to the environmental criterion, measured as a correlation between Y_e and $\hat{Y}_e$. Similarly, R_s represents cognitive control in that it is a measure of how well the human judgment model matches the actual human judgment (the correlation between Y_s and $\hat{Y}_s$). Finally, C represents unmodeled agreement: a measure of the correspondence between the information not captured between the two models. This is measured as the correlation between the residuals of the environment model ($\hat{Y}_e - Y_e$) and the judgment model ($\hat{Y}_s - Y_s$).

JA affords numerous analysis capabilities, including performance and judgment policy typing (Cooksey, 1996). Judgment performance typing groups judges based on the similarity of their lens model statistics. Groupings are often based on policy consistency, or cognitive control (R_s), linear knowledge (G), or r_m (the linear knowledge equivalent when comparing two judges to each other instead of a judge to the environment) (Cooksey, 1996). A judgment policy is defined by which cues are considered and how they are combined to make a judgment (Bisantz and Pritchett, 2003). Judgment policy typing describes the grouping of judges based on the similarity of cue weighting.

In this form, the double system lens model has been successfully used to model and evaluate human judgment in a number of domains including policy making (Dalgleish, 1988; Hammond, 1996), medicine (Wigton, 1988), weather forecasting (Stewart et al., 1992), education (Cooksey and Freebody, 1987), air traffic control (Bass and Pritchett, 2008; Bisantz and Pritchett, 2003), and many other (Karelaia and Hogarth, 2008; Kaufmann et al., 2013). To date, it has never been used to evaluate phishing judgments.

Because this form of judgment analysis is based on multiple linear regression, it is important that the assumptions of this statistical approach are maintained in the resulting models. Most important to the double system lens model is that cues should be linearly combinable in a way that is meaningful for predicting both the criterion and the human judgment. This is not always the case as environmental cues may not be linearly related to the criterion, and there may not be clear transformations to account for these nonlinearities. Further, human judgments can be nonlinear when they are under time pressure, using intuition or imagination, considering multiple alternatives, or matching patterns from their experience to the current situation (Hogarth, 2001).

2.3. *Phishing cues*

Beyond the linearity considerations, cue identification can be a major challenge for many lens model analyses (Cooksey, 1996). Work on automated phishing detection methods have

Table 1 – Phishing cues.

Cue category	Cue name	Description
Technical	Sender display name and email address	Display names are easily spoofed and can hide the sender's real email address (Blythe et al., 2011; Downs et al., 2006; Furnell, 2007; Karakasiliotis et al., 2006; Kim and Hyun Kim, 2013; Vishwanath, 2016; Vishwanath et al., 2011; Wang et al., 2012).
	URL hyperlinking*	URL hyperlinking hides the true URL behind text; the text can also look like another link (Canfield et al., 2016; Downs et al., 2006; Egelman et al., 2008; Furnell, 2007; Jakobsson, 2007; Karakasiliotis et al., 2006).
	Attachment type	The presence of file attachments, especially an executable, can be a phishing indicator (Han and Shen, 2016).
Visual presentation	No branding/logos*	No or very minimal branding and logos can be a sign of a suspicious email (Blythe et al., 2011; Furnell, 2007; Grazioli, 2004; Jakobsson, 2007; Karakasiliotis et al., 2006; Kim and Hyun Kim, 2013; Tsow and Jakobsson, 2007; Vishwanath, 2016).
	Poor overall design/formatting	Generally poor formatting and design or an overall unprofessional look can be a sign of a suspicious email (Dhamija et al., 2006; Fogg, 2003; Jakobsson, 2007; Karakasiliotis et al., 2006; Parsons et al., 2013; Tsow and Jakobsson, 2007).
Message language and content	Spelling and grammar errors*	Emails with multiple spelling or grammar errors can be suspicious (Blythe et al., 2011; Canfield et al., 2016; Downs et al., 2006; Furnell, 2007; Grazioli, 2004; Jakobsson, 2007; Jakobsson and Finn, 2007; Karakasiliotis et al., 2006; Parsons et al., 2013; Vishwanath et al., 2011; Wang et al., 2012; Wright et al., 2010).
	Generic greeting*	A generic greeting and an overall lack of personalization in the email can be an indicator of a suspicious email (Alsharnouby et al., 2015; Canfield et al., 2016; Downs et al., 2006; Egelman et al., 2008; Karakasiliotis et al., 2006; Parsons et al., 2013; Tsow and Jakobsson, 2007).
	Use of time pressure/threatening language*	Phishing emails often use time pressure or threats (ex. legal ramifications) to try to get users to quickly comply with the request (Alsharnouby et al., 2015; Canfield et al., 2016; Downs et al., 2006; Karakasiliotis et al., 2006; Kim and Hyun Kim, 2013; Vishwanath et al., 2011).
	Use of emotional appeals	Phishers can try to appeal to a user's emotions with humanitarian claims (ex. donating money to the poor) (Karakasiliotis et al., 2006; Kim and Hyun Kim, 2013).
	Lack of signer details*	Emails including few details about the sender, like contact information, can be suspicious (Kim and Hyun Kim, 2013).
	Too good to be true offers*	Emails offering contest winnings or other unlikely monetary and/or material benefits can be suspicious (Grazioli, 2004; Karakasiliotis et al., 2006; Parsons et al., 2013; Wright et al., 2010).
	Requests for personal information*	Requests for personal information, like a social security number, can indicate a suspicious email (Downs et al., 2006; Furnell, 2007).

Note. Cues marked with an asterisk were included in the analysis presented in this paper.

identified a number of different indicators (candidate cues) that can provide evidence that an email is phishing. These potential cues are generally divided into three main categories (Karakasiliotis et al., 2006): technical, visual presentation, and message language and content. Cue types, names, and descriptions are summarized in Table 1.

Assuming these cues are not highly correlated, each can provide additional evidence about whether an email is phishing. This suggests that cues should be linearly combinable in a JA context.

2.4. *Objective*

The double system lens model offers powerful means of evaluating human judgments. As such, we would like to be able to use these to analyze phishing judgments. However, to be able to do this, the phishing cues should linearly combine together to help predict both human judgments and the criterion. Given that each of the phishing cues should provide different information about whether an email is or is not a phishing attempt, we hypothesize that the phishing cues are linearly combinable and thus that the double system lens model is appropriate for analyzing phishing. In the presented work, we aimed to test this hypothesis. To do this, we used an existing data set in which humans made judgments about whether emails were phishing or not. We used the double system lens model to analyze this data set to determine how well the lens model captured the judgment task.

3. Methods

3.1. *Experimental task*

The data used in the research reported here were collected for an independent effort. However, they contained the information necessary for the presented analysis. In the experimental task, participants were told that they were an administrative assistant and that their boss, department chair Dr. Jane Smith, asked them to sort through her emails while she was on vacation. Participants were told that the chair uses her email for many different accounts, both work and personal. Participants did not need to respond to any of the emails, only sort them into either a "keep" or "suspicious" folder. Participants were

also asked not to use the internet or other sources to look up anything about the emails. Their judgment of the email should only be based on the information within the email and email client. While following a think aloud protocol, participants had 30 min to sort 40 emails (20 legitimate and 20 phishing). Participants were instructed to prioritize explaining their decision making process and doing the task correctly over doing the task quickly.

3.2. Participants

Ten students participated in the study. Participants averaged 23.2 years of age. Six were male and four were female. Five were native English speakers. Participants had varying levels of cybersecurity knowledge. All participants had a basic understanding of what phishing emails are and strategies for dealing with them. One participant was a previous systems administrator. All participants reported that they spent multiple hours a day working on a computer and regularly checked their email using both a computer and a smartphone. Participants were run one at a time. The population from which participants were recruited was large enough to make it unlikely that interactions between participants could have influenced the results.

3.3. Apparatus

The experiment was conducted in a controlled office environment on personal computers (PC) and mobile smartphones (Mobile). These platforms allowed participants to interact with Roundcube, a web-based email client with a skin resembling that of Microsoft Outlook. This allowed for natural email tasks and included all standard email functionality, like hovering over links and the sender's display name and moving emails into different folders.

Screen recording software and human observers were used to collect dependent measure data from participants.

3.4. Independent variables and experimental design

Participants were randomly assigned to a technology condition, either PC or Mobile. For the PC condition, participants completed the email sorting task on a desktop computer. For the mobile condition, participants used a smartphone to complete the task. Five participants were in each technology condition. As previously noted, these data were collected as part of a separate effort. Thus, while we were not specifically interested in the effect of the technology condition in this research, it was part of the original experimental design.

The criterion for each email considered by participants was coded as a dichotomous variable, where an email was either a phishing email (coded as 1) or not (coded as 0). After the cue list was set, double system lens model analyses (see Section 2.2) were conducted for all 10 participants. Cues were also coded as dichotomous variables for each email inspected by participants, where a 1 meant the cue was present and 0 meant it was not. This included the presence of URL hyperlinking, the presence of attachments, whether branding or logos were absent, whether there were spelling and/or grammar errors, whether there was a generic greeting, whether time pressure or threatening language was used, whether there was a lack of signer details, if they had too good to be true offers, and if they requested for personal information. Whether or not emotional appeals were in the message was coded, but no emails included this cue. Thus, this cue was not considered in the presented analyses. Because participants were not evaluating their own inbox, sender display name and email address was not an appropriate cue in the context of the experimental task; it was not included. Because the community does not have a standardized method to assess poor overall design/formatting, this cue was also not included.

Participants were given 40 emails to sort, presented in a random order for each participant. All participants were presented with the same 40 emails. Twenty were legitimate and 20 were phishing; participants were not aware of this distribution. All emails were created from real emails (either phishing or legitimate) with personal identifying information modified from the original sender and recipient to prevent the distribution of any personal information. Phishing emails were derived from a semi-random sample of emails in Cornell University's "Phish Bowl" phishing email database (Cornell University, 2017). Legitimate emails were derived from legitimate emails received by the researchers that conducted the study.

3.5. Dependent measures

Dependent measures included the judgment the participants made about an email: 1 if the participant thought it was phishing (moved the email to the suspicious folder) and 0 if not (moved the email to the keep folder). The time spent on each email, the number of times the email header was opened, the number of times the sender was hovered over (or the mobile equivalent), and the number of times a link was hovered over (or the mobile equivalent) were also collected (though not considered in the presented analyses).

3.6. Data analysis

There were 40 original emails. Ultimately, because cues relevant for attachment-based emails are not the same as link-based ones, we limited our analysis to link-based phishing emails. Thus, the attachment-based emails were removed, resulting in a total of 38 emails: 20 legitimate and 18 phishing. Removing attachment-based emails also resulted in the attachment type cue being excluded from consideration in the analysis.

Double system lens model analyses were completed with the Cognitive Systems Engineering Educational Software (CSEES) (Bolton and Bass, 2005). The lens model analyses commenced by checking for inter-cue correlations to ensure that our analysis did not include redundant cues. In cases where cues were highly correlated (a Pearson's correlation coefficient was significant, $p < 0.05$), only one of the correlated cues was considered in the final analysis. This resulted in the inclusion of eight cues: spelling and grammar errors, generic greeting, URL hyperlinking, no branding/logos, lack of signer details, too good to be true offers, requests for personal information, and use of time pressure/threatening language. The number of times each cue was present across the 38 emails included in the analysis, as well as the number and types of cues in each

		Presence of Cue								
	Email	Spelling	Greeting	Linked	Branding	Personal	Signer	TooGood	Pressure	**Total**
Legitimate	1	0	1	1	0	0	0	0	0	**2**
	2	0	1	1	0	0	0	0	0	**2**
	3	0	1	0	1	0	0	0	0	**2**
	4	0	1	1	0	0	0	0	0	**2**
	5	0	0	1	1	0	0	0	0	**2**
	6	0	1	1	0	0	0	0	0	**2**
	7	0	1	1	0	0	0	0	0	**2**
	8	0	1	1	0	0	1	0	0	**3**
	9	0	1	1	0	0	0	0	0	**2**
	10	0	1	1	0	0	0	0	0	**2**
	11	0	1	1	0	0	0	0	0	**2**
	12	0	0	1	0	0	0	0	0	**1**
	13	0	1	1	0	0	0	0	0	**2**
	14	0	0	1	1	0	0	0	0	**2**
	15	0	0	0	1	1	0	0	0	**2**
	16	0	1	1	0	0	0	0	0	**2**
	17	0	1	0	1	0	0	0	0	**2**
	18	0	0	1	0	0	0	0	1	**2**
	19	1	1	0	1	0	0	0	0	**3**
	20	0	0	0	1	0	0	0	0	**1**
	Total	**1**	**14**	**15**	**7**	**1**	**1**	**0**	**1**	
Phishing	1	0	1	1	0	0	1	0	1	**4**
	2	0	0	1	1	1	1	0	0	**4**
	3	0	1	1	1	0	0	0	0	**3**
	4	0	1	1	1	0	1	0	0	**4**
	5	1	1	1	0	0	1	1	1	**6**
	6	1	1	1	1	1	1	0	0	**6**
	7	1	1	1	1	0	1	0	0	**5**
	8	0	1	1	0	0	1	0	1	**4**
	9	1	1	0	1	0	1	0	1	**5**
	10	0	1	1	1	0	1	0	0	**4**
	11	1	1	1	1	0	1	0	0	**5**
	12	0	1	1	0	1	1	0	1	**5**
	13	1	1	1	0	0	1	0	0	**4**
	14	1	1	1	1	1	1	0	1	**7**
	15	1	1	1	1	1	1	0	1	**7**
	16	1	1	1	1	0	0	0	1	**5**
	17	1	1	1	0	1	1	1	0	**6**
	18	1	1	1	1	1	1	0	0	**6**
	Total	**11**	**17**	**17**	**12**	**7**	**16**	**2**	**8**	

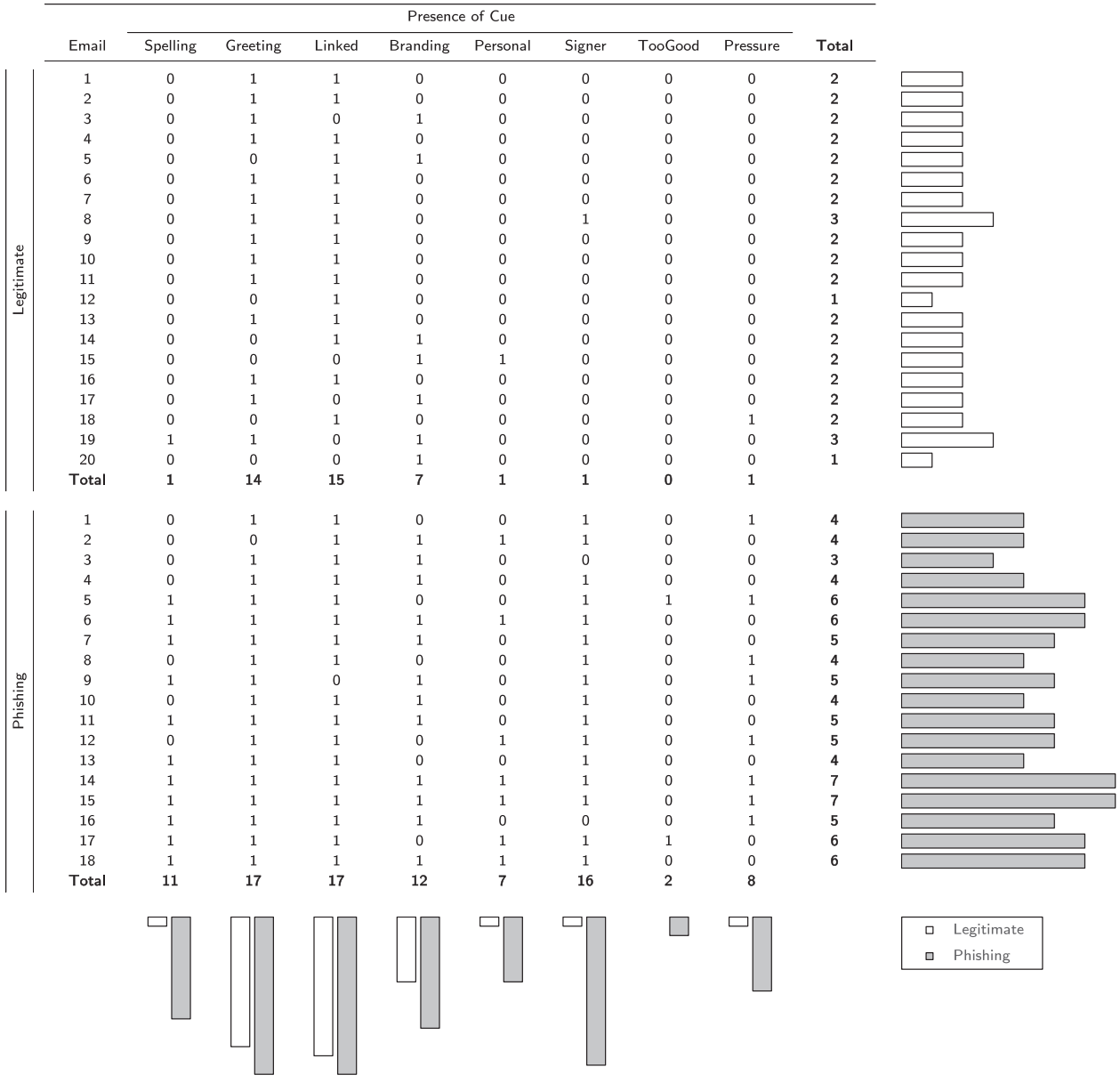

Fig. 2 – Type and number of cue occurrences for phishing and legitimate emails.

email, is shown in Fig. 2. On average, there were more cues present in the phishing emails ($M = 5$, $SD = 1.14$) compared to the legitimate emails ($M = 2$, $SD = 0.46$).

To determine if significant differences manifested between the PC and mobile conditions, t-tests were used to compare average sorting accuracy and judgment strategy (cue β weights). A Bonferroni adjustment was used to account for multiple comparisons (Dunn, 1961).

4. Results

The results of the double system lens model analyses (cue weights and statistics) are presented in Table 2. These show that there was a range of achievement values (r_a) across participants, meaning overall task performance varied. The large R_e (0.923) and R_s values (ranging from the maximum possible value of 1 down to 0.783) indicate that the multiple linear regression models did a good job of fitting both the environment and the human judges. G values (linear knowledge) were also high across the board, indicating that the linear regression models of the human judges generally matched the linear model of the environment. Conversely, a large range of C values (unmodeled agreement) were observed. This suggests that there are distinct individual differences between participants.

The beta weight of each cue in the criterion model (ecological validity) was positive. This indicates that each cue provided some evidence about whether the email was phishing or not. Furthermore, because each cue was coded on the same scale (present or not present) we can compare their relative weights to understand how important each was for determining if an email was phishing. For example, lack of signer details (0.542), no branding/logos (0.400), and URL hyperlinking (0.349)

Table 2 – Regression and lens model analyses results.

Criterion	Regression models									Lens model statistics			
Participant	β_0	$\beta_{Spelling}$	$\beta_{Greeting}$	β_{Linked}	$\beta_{Branding}$	$\beta_{Personal}$	β_{Signer}	$\beta_{TooGood}$	$\beta_{Pressure}$		R_e		
	−0.510	0.033	0.194	0.349	0.400	0.007	0.542	0.255	0.268	r_a	0.923 R_s	G	C
1	0.175	0.475	−0.009	−0.106	0.040	0.309	0.223	−0.045	0.262	0.789	0.840	0.834	0.679
2	−0.226	0.141	0.265	0.322	0.106	−0.225	0.387	−0.272	0.383	0.669	0.771	0.897	0.138
3	−0.038	0.556	−0.060	0.101	−0.065	0.341	0.301	−0.584	0.105	0.760	0.877	0.834	0.460
4	0.127	0.419	0.024	−0.129	−0.087	−0.121	0.490	0.087	−0.018	0.709	0.792	0.861	0.342
5	−0.510	0.033	0.194	0.349	0.400	0.007	0.542	0.255	0.268	∼1.000	0.923	∼1.000	∼1.000
6	−0.216	0.077	0.060	0.154	0.172	0.012	0.632	−0.326	0.224	0.851	0.841	0.960	0.511
7	−0.510	0.033	0.194	0.349	0.400	0.007	0.542	0.255	0.268	∼1.000	0.923	∼1.000	∼1.000
8	−0.118	−0.018	0.201	−0.048	0.218	0.341	0.310	0.011	−0.016	0.489	0.648	0.842	−0.049
9	0.018	0.108	0.117	0.098	0.060	0.234	0.461	0.003	0.156	0.748	0.749	0.952	0.350
10	0.263	0.065	0.109	−0.084	−0.088	−0.363	0.642	0.136	0.102	0.527	0.676	0.783	0.134

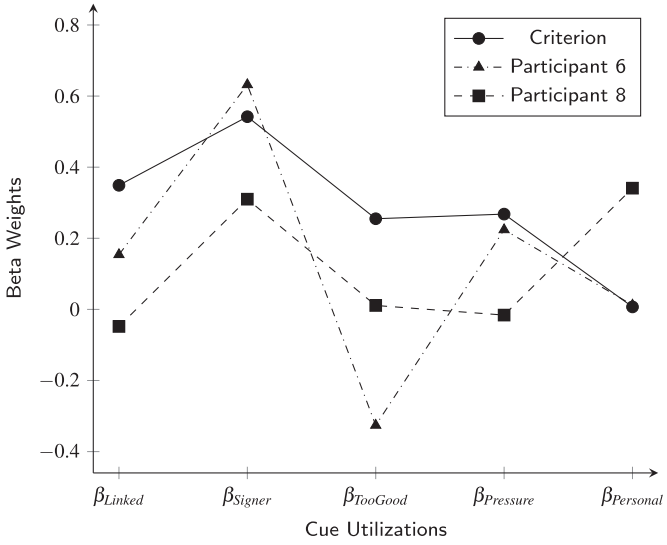

Fig. 3 – Comparison of criterion, participant 6, and participant 8 beta weights for a subset of cues.

appear to be the most diagnostic, while requests for personal information (0.007), spelling and grammar errors (0.033), and generic greeting (0.194) were the least.

Similarly, we can compare judgment strategies between participants and between given participants and the environment model. When looking at the participants with the highest achievement values (r_a), participants 5 and 7 had cue utilizations that effectively matched the corresponding ecological validities in the environment (criterion) model. Conversely, cue utilizations from the participant with the lowest achievement, participant 8, differed greatly from the cue validities in the criterion model. In particular, participant 8 (Fig. 3) appeared to under-weight URL hyperlinking, lack of signer details, too good to be true offers, and use of time pressure/threatening language while over-weighting requests for personal information. The cue utilizations for the participant with the second highest achievement, participant 6, do not perfectly match the criterion (Fig. 3). However, this appears to predominantly occur because of the under-weighting of too good to be true offers.

Table 3 – Comparison of experimental conditions.

	Mobile	PC	*p*-value	Bonferroni adj.
Sorting accuracy	0.844	0.900	0.335	1.000
$\beta_{Spelling}$	0.139	0.239	0.487	1.000
$\beta_{Greeting}$	0.152	0.067	0.219	1.000
β_{Linked}	0.087	0.115	0.831	1.000
$\beta_{Branding}$	0.135	0.096	0.756	1.000
$\beta_{Personal}$	0.014	0.094	0.634	1.000
β_{Signer}	0.421	0.485	0.510	1.000
$\beta_{TooGood}$	0.017	−0.113	0.480	1.000
$\beta_{Pressure}$	0.200	0.147	0.557	1.000

As shown in Table 3, no significant differences were observed between in sorting accuracy or cue weights between the mobile and PC technology conditions.

5. Discussion and conclusion

The double system lens model gives analysts powerful tools for evaluating and understanding human judgments. The pre-

sented work is the first step in applying JA to the phishing domain. Overall, the results indicate that the lens model can be used as a means for evaluating phishing. The best evidence for this is seen in the high R_e and R_s values we observed. The high R_e value suggests that linear regression is an appropriate method for predicting whether or not an email is phishing based on cues available to the human operator. The high R_s values indicate that the linear regression model is able to adequately capture the human's judgment strategy so that it can be compared to that of the criterion and other human judges.

The lens model's compatibility with this domain affords many additional analysis capabilities. First, analysts can compare cue validities and cue utilizations. This allows for the investigation of whether or not there is a mismatch between what cues are the most diagnostic and the cues actually being used by human operators. In related work, Parsons et al. (2016) used a series of t-tests to determine which cues (independently of others) best differentiate phishing from legitimate emails and which cues are used by human judges. As our results show, the JA approach not only allows the assessment of the relative importance of different cues, but also accounts for how all of the included cues combine. The JA approach also enables us to compare the judgment strategies of different humans to each other and to the environment model. This was shown in the differences we saw in judgment strategy between participants based on their cue utilizations. From an engineering standpoint, this has the potential to help poor performers because it could be used to inform or train humans to modify their judgment strategy to account for under or over-weightings. Thus, the JA approach capture more nuanced aspects of the phishing judgment task and deeper analyses capabilities than the previous study (Parsons et al., 2016).

It is important to note that, while our results are compelling, we cannot make any claims about their generalizability due to dataset limitations. The distribution of emails and phishing cues in the experiment is consistent with previous phishing research (Canfield et al., 2016; Dhamija et al., 2006; Kumaraguru et al., 2010; Pattinson et al., 2012). However, we acknowledge this distribution (Fig. 2) may not be realistic. The dataset also included PC and Mobile experimental conditions and there is little research to understand how this may influence cue processing. Although our results did not indicate any significant differences between groups, Vishwanath (2016) noted that mobile device usage strengthened email habits, potentially resulting in an increased likelihood of victimization. It is important to note that our results comparing the two technology conditions groups should be interpreted cautiously because of the small sample size. Future work should develop experimental methods to improve generalizability and applicability across computing platforms.

Further note that the cues, judgments, and criterion used in this study are all dichotomous variables. Ideally, we would have used logistic regression in our JA instead of multiple linear regression, because it is more appropriate for handling dichotomous dependent measures. However, our experimental design did not allow for this due to logistic regression's vulnerability to inter-cue correlations in small data sets (Hamm and Yang, 2017). Despite this limitation, our use of multiple linear regression is acceptable. Specifically, Cooksey (1996, p. 298) noted that the use of multiple linear regression for dichotomous variables is "acceptable where the researcher is chiefly interested in the Lens Model correlations and cue weights, and is not interested in examining the precise predicted values arising from each regression model," which is consistent with the way we interpreted our results. However, to fully realize the power of JA for modeling phishing judgments, future work should investigate how logistic regression and its associated lens model equation can be applied to this domain (Stewart, 2004).

Finally, while the sample size used in our study was small, this is not a limitation for the primary purpose of our analyses. Specifically, double system lens model analyses are, by definition, designed to evaluate the judgments of individuals. Thus, the fact that the lens model analyses produced compelling results for all of the analyzed participants provided us with ample evidence to test our hypothesis. However, the lens model does support analyses for clustering humans based on their judgment strategies and statistically comparing lens model parameters between individuals and groups (Cooksey, 1996). To allow such capabilities to be utilized in analyses, future work should employ larger numbers of participants.

Acknowledgments

The authors would like to thank Dr. Anton Dahbura and Dr. Xiangyang Li from the Johns Hopkins University Information Security Institute and Dr. Nathan Bos from the Johns Hopkins University Applied Physics Laboratory for allowing them to use the data collected under the National Science Foundation Award 1544493 for the work presented here.

REFERENCES

Alsharnouby M, Alaca F, Chiasson S. Why phishing still works: user strategies for combating phishing attacks. Int J Hum-Comput Stud 2015;82:69–82.

Anti-Phishing Working Group. Phishing activity trends report 4th quarter 2016. Anti-Phishing Working Group 2017.

Arachchilage NAG, Love S. Security awareness of computer users: a phishing threat avoidance perspective. Comput Hum Behav 2014;38:304–12.

Bass EJ, Pritchett AR. Human-automated judge learning: a methodology for examining human interaction with information analysis automation. IEEE Trans Syst Man Cybern Part A Syst Hum 2008;38(4):759–76.

Bisantz AM, Pritchett AR. Measuring the fit between human judgments and automated alerting algorithms: a study of collision detection. Hum Factors J Hum Factors Ergon Soc 2003;45(2):266–80.

Blythe M, Petrie H, Clark JA. F for fake: four studies on how we fall for phish. Proceedings of the 2011 SIGCHI conference on human factors in computing systems. ACM; 2011. p. 3469–78.

Bolton ML, Bass EJ. Cognitive systems engineering educational software (CSEES): educational software addressing quantitative models of performance. Proceedings of the 2005 IEEE international conference on systems man and cybernetics. IEEE; 2005. p. 3380–6.

Bruins HH, Cooksey RW. JANNET, a neural network approach to judgment analysis. Preradiation dental decisions in patients with head and neck cancer. Utrecht: University Medical

Center Utrecht, 2000. http://dspace.library.uu.nl/bitstream/handle/1874/393/c4.pdf.
Brunswik E. Representative design and probabilistic theory in a functional psychology. Psychol Rev 1955;62(3):193.
Canfield CI, Fischhoff B, Davis A. Quantifying phishing susceptibility for detection and behavior decisions. Hum Factors J Hum Factors Ergon Soc 2016;58(8):1158–72.
Caputo DD, Pfleeger SL, Freeman JD, Johnson ME. Going spear phishing: exploring embedded training and awareness. IEEE Secur Priv 2014;12(1):28–38.
Cooksey RW. Judgment analysis: Theory, methods, and applications. Academic Press; 1996.
Cooksey RW, Freebody P. Cue subset contributions in the hierarchical multivariate lens model: judgments of children's reading achievement. Organ Behav Hum Decis Process 1987;39(1):115–32.
Cornell University. Phish bowl. Cornell University 2017.
Dalgleish LI. Decision making in child abuse cases: applications of social judgment theory and signal detection theory. Adv Psychol 1988;54:317–60.
Davinson N, Sillence E. It wont happen to me: promoting secure behaviour among internet users. Comput Hum Behav 2010;26(6):1739–47.
Dhamija R, Tygar JD, Hearst M. Why phishing works. Proceedings of the SIGCHI conference on human factors in computing systems. ACM; 2006. p. 581–90.
Downs JS, Holbrook MB, Cranor LF. Decision strategies and susceptibility to phishing. Proceedings of the second symposium on usable privacy and security. ACM; 2006. p. 79–90.
Dunn OJ. Multiple comparisons among means. J Am Stat Assoc 1961;56(293):52–64.
Egelman S, Cranor LF, Hong J. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. Proceedings of the SIGCHI conference on human factors in computing systems. ACM; 2008. p. 1065–74.
Ferguson AJ. Fostering e-mail security awareness: the west point carronade. Educase Quarterly 2005;28(1):54–7.
Fogg BJ. Prominence-interpretation theory: explaining how people assess credibility online. Proceedings of the CHI'03 extended abstracts on human factors in computing systems. ACM; 2003. p. 722–3.
Furnell S. Phishing: can we spot the signs? Comput Fraud Secur 2007;2007(3):10–15.
Grazioli S. Where did they go wrong? An analysis of the failure of knowledgeable internet consumers to detect deception over the internet. Group Decis Negot 2004;13(2):149–72.
Hamm RM, Yang H. Alternative lens model equations for dichotomous judgments about dichotomous criteria. J Behav Decis Mak 2017;30(2):527–32.
Hammond KR. Human judgement and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice. Oxford University Press; 1996.
Han Y, Shen Y. Accurate spear phishing campaign attribution and early detection. Proceedings of the thirty-first annual ACM symposium on applied computing. ACM; 2016. p. 2079–86.
Hogarth RM. Educating intuition. Chicago: University of Chicago Press; 2001.
Hong J. The state of phishing attacks. Commun ACM 2012;55(1):74–81.
Hursch CJ, Hammond KR, Hursch JL. Some methodological considerations in multiple-cue probability studies. Psychol Rev 1964;71(1):42.
Jakobsson M. The human factor in phishing. Priv Secur Consum Inf 2007;7(1):1–19.
Jakobsson M, Finn P. Designing and conducting phishing experiments. IEEE Technol Soc Mag 2007;26(1):46–58. Special Issue on Usability and Security.
Karakasiliotis A, Furnell S, Papadaki M. Assessing end-user awareness of social engineering and phishing. Proceedings of the 2006 Australian information warfare and security conference. School of Computer and Information Science, Edith Cowan University, Perth, Western Australia, 2006.
Karelaia N, Hogarth RM. Determinants of linear judgment: a meta-analysis of lens model studies. Psychol Bull 2008;134(3):404–26.
Kaufmann E, Reips UD, Wittmann WW. A critical meta-analysis of lens model studies in human judgment and decision-making. PLOS One 2013;8(12):1–16.
Kim D, Hyun Kim J. Understanding persuasive elements in phishing e-mails: a categorical content and semantic network analysis. Online Inf Rev 2013;37(6):835–50.
Krebs B, FBI: 2.3 billion lost to ceo email scams, 2016, https://krebsonsecurity.com/2016/04/fbi-2-3-billion-lost-to-ceo-email-scams/.
Kumaraguru P, Rhee Y, Acquisti A, Cranor LF, Hong J, Nunge E. Protecting people from phishing: the design and evaluation of an embedded training email system. Proceedings of the SIGCHI conference on human factors in computing systems. ACM; 2007. p. 905–14.
Kumaraguru P, Sheng S, Acquisti A, Cranor LF, Hong J. Lessons from a real world evaluation of anti-phishing training. Proceedings of the 2008 eCrime researchers summit. IEEE; 2008. p. 1–12.
Kumaraguru P, Sheng S, Acquisti A, Cranor LF, Hong J. Teaching johnny not to fall for phish. ACM Trans Internet Technol (TOIT) 2010;10(2):7.
Microsoft. Technical report. 2013 microsoft computing safety index (MCSI) worldwide results summary. Microsoft Corporation; 2014.
Parsons K, Butavicius M, Pattinson M, Calic D, Mccormac A, Jerram C. Do users focus on the correct cues to differentiate between phishing and genuine emails?. Proceedings of the 2015 Australasian conference on information systems (ACIS), 2016.
Parsons K, McCormac A, Pattinson M, Butavicius M, Jerram C. Phishing for the truth: a scenario-based experiment of users behavioural response to emails. Proceedings of the 2013 IFIP international information security conference. Springer; 2013. p. 366–78.
Passeri P, 2015 cyber attacks statistics, 2016, http://hackmageddon.com/2016/01/11/2015-cyber-attacks-statistics/.
Pattinson M, Jerram C, Parsons K, McCormac A, Butavicius M. Why do some people manage phishing e-mails better than others? Inf Manag Comput Secur 2012;20(1):18–28.
Pfleeger SL, Caputo DD. Leveraging behavioral science to mitigate cyber security risk. Comput Secur 2012;31(4):597–611.
Stewart T. Notes on a form of the lens model equation for logistic regression analysis. Proceedings of the 2004 Brunswik society meeting, 2004.
Stewart TR, Heideman KF, Moninger WR, Reagan-Cirincione P. Effects of improved information on the components of skill in weather forecasting. Organ Behav Hum Dec Process 1992;53(2):107–34.
Tsow A, Jakobsson M, 9. Indiana University; 2007.
Tucker LR. A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. Psychol Rev 1964;71(6):528.
Verizon. 2017 data breach investigations report 10th edition. Verizon 2017.
Vishwanath A. Mobile device affordance: explicating how smartphones influence the outcome of phishing attacks. Comput Hum Behav 2016;63:198–207.

Vishwanath A, Harrison B, Ng YJ. Suspicion, cognition, and automaticity model of phishing susceptibility. Commun Res 2016.
Vishwanath A, Herath T, Chen R, Wang J, Rao HR. Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. Decis Support Syst 2011;51(3):576–86.
Volz D, Number of U.S. government 'cyber incidents' jumps in 2015, 2016, http://www.reuters.com/article/us-usa-cyber-iduskcn0wn263.
Wang J, Herath T, Chen R, Vishwanath A, Rao HR. Phishing susceptibility: an investigation into the processing of a targeted spear phishing email. IEEE Trans Prof Commun. 2012;55(4):345–62.
Wigton RS. Applications of judgment analysis and cognitive feedback to medicine. Adv Psychol 1988;54:227–45.
Williams EJ, Beardmore A, Joinson AN. Individual differences in susceptibility to online influence: a theoretical review. Comput Hum Behav 2017;72:412–21.
Wright R, Chakraborty S, Basoglu A, Marett K. Where did they go right? Understanding the deception in phishing communications. Group Decis Negot 2010;19(4):391–416.
Yoon JM, He D, Bolton ML. A lamstar network-based human judgment analysis. IEEE Trans Hum-Mach Syst 2017;47(6):951–7.
Zielinska OA, Welk AK, Mayhorn CB, Murphy-Hill E. Exploring expert and novice mental models of phishing. Proceedings of the 2015 human factors and ergonomics society annual meeting. SAGE Publications; 2015. p. 1132–6.

Kylie A. Molinaro received the B.S. degree in human factors psychology in 2014 from Embry-Riddle Aeronautical University and the M.S. degree in industrial engineering in 2016 from the University at Buffalo, The State University of New York. She is currently working toward the Ph.D. degree in industrial engineering in the Department of Industrial and Systems Engineering, University at Buffalo, The State University of New York. She is a Human Factors Engineer with the Johns Hopkins University Applied Physics Laboratory. Her research focuses on judgment and decision-making analyses in the cybersecurity domain.

Matthew L. Bolton received the B.S. in computer science in 2004, the M.S. in systems engineering in 2006, and the Ph.D. in systems engineering in 2010 from the University of Virginia, Charlottesville, USA. He is currently an Assistant Professor in the Department of Industrial and Systems Engineering at the University at Buffalo, the State University of New York. His research focuses on the use of human performance modeling and formal methods in the analysis, design, and evaluation of safety-critical systems and cybersecurity.