Computers in Human Behavior 87 (2018) 174–182

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Full length article

Phishing attempts among the dark triad: Patterns of attack and vulnerability

Shelby R. Curtis[a], Prashanth Rajivan[b], Daniel N. Jones[a], Cleotilde Gonzalez[b,*]

[a] University of Texas at El Paso, El Paso, USA
[b] Carnegie Mellon University, Pittsburgh, PA, 15213, USA

ARTICLE INFO

Keywords:
Phishing
Deception
Human behavior
Personality
Dark triad

ABSTRACT

Phishing attacks are more common and more sophisticated than other forms of social engineering attacks. This study presents an investigation of the relationships between three personality traits—Machiavellianism, narcissism, and psychopathy (i.e., the Dark Triad)—and phishing effort, attack success, and end-user susceptibility to phishing emails. Participants were recruited in two stages. The first set of participants acted as attackers, creating phishing emails. The second set of participants acted as end-users, reading both benevolent and phishing emails and indicating their likely behavioral response to each email. Our findings suggest that attackers' Dark Triad scores relate to the effort that they put in writing a phishing email, but do not predict phishing success. Instead, it is the end-users' Dark Triad scores that predict the success of phishing emails. We found that higher levels of attacker Machiavellianism were linked to increased phishing effort, while end-user narcissism was associated to greater vulnerability when receiving phishing emails. Furthermore, our findings suggest that narcissistic end-users were marginally more susceptible to phishing emails that originated from narcissistic attackers. These results have important practical implications for training, anti-phishing tool development, and policy in organizations.

1. Introduction

If you received an email titled, “Urgent! You need to verify recent account activity!” would you open it? Many people would. Phishing is the most common form of cyberattack in which criminals (attackers) deceive people via socially-engineered strategies into installing harmful software or surrendering sensitive information (Anderson, 2010; Hong, 2012). Although technologies such as spam filters have been developed to effectively detect and deter known phishing campaigns, attackers continuously find new ways to evade these technologies such as through sophisticated and personalized e-mails (“spear phishing”) that take advantage of human limitations and biases and persuade people to respond (Im & Baskerville, 2005).

Past work on human behavior in phishing attacks has investigated attackers' approaches, perspectives, and attitudes toward system abuse (Willison & Backhouse, 2006) and end-users’ responses to malicious emails (e.g., Stanton, Stam, Mastrangelo, & Jolton, 2005). However, few researchers have addressed how individual differences in personality traits relate to attack strategies and end-user vulnerabilities. For example, Cho, Cam, and Oltramari (2016) investigated the correlation between personality traits and end-user vulnerability to phishing attacks and found that agreeableness and neuroticism were positively correlated with perceived trust and risk taking. However, little is known regarding attacker personality and behavior, and research has not yet examined how relationships between the personalities of both attacker and end-user may contribute to phishing success.

A recent study spearheaded the analysis of human adversarial behavior in phishing (Rajivan & Gonzalez, 2018). In an experiment, participants were asked to design phishing emails and were rewarded for their ability to evade detection and persuade end-users to respond. The reward function and the timing of the reward were experimentally controlled. In a second phase of the experiment, an equal mix of these phishing emails and ham (benign emails) were distributed to end-users to decide how they would respond to these emails. Importantly, this study measured Machiavellianism, narcissism, and psychopathy for both, attackers and end-users, using the Short Dark Triad questionnaire (see below, Jones & Paulhus, 2014).

In the present research, we take up this mantle and revisit the data from Rajivan and Gonzalez (2018) to determine how attacker and end-user personalities relate to phishing success. In what follows, we develop predictions based on the Dark Triad literature, present our methods for the analyses, report our findings, and discuss their implications for phishing prevention and awareness.

* Corresponding author. Social and Decision Sciences Department Carnegie Mellon University, 5000 Forbes Ave. Porter Hall 208 Pittsburgh, PA, 15213 USA..
E-mail address: coty@cmu.edu (C. Gonzalez).
URL: https://www.cmu.edu/ddmlab (C. Gonzalez).

https://doi.org/10.1016/j.chb.2018.05.037
Received 29 March 2018; Received in revised form 23 May 2018; Accepted 28 May 2018
Available online 31 May 2018

1.1. The dark triad

Machiavellianism, narcissism, and psychopathy, together known as the “Dark Triad” of personality traits (Paulhus & Williams, 2002), have been studied in the context of interpersonal manipulation and deception across a variety of situations (Jonason & Webster, 2012). Machiavellianism is associated with manipulative behavior aimed at maximizing personal gain through strategic deception and flexible moral tactics (Bereczkei, 2015; Christie & Geis, 1970). Narcissism is associated with interpersonal dominance, entitlement, and the willingness to exploit others (Emmons, 1987; McHoskey, 1995). In zero-acquaintance paradigms where people meet for the first time in social interactions, people high in narcissism perform especially well: they are perceived as more popular, more appealing, and more likeable (Back, Schmukle, & Egloff, 2010; Jauk et al., 2016). Psychopathy is associated with the absence of empathy and with tendencies toward impulsivity, aggression, and deception (Azizli et al., 2016; Williams, Paulhus, & Hare, 2007), which lead to reckless behavior and the successful imitation of socially appropriate emotions and intentions in short-term encounters (Book et al., 2015).

Traditionally, the Dark Triad has been assessed using popular instruments that were developed to measure each of the three dark personality traits in isolation (McHoskey, Szarzto, & Worzel, 1998). Currently, there are two brief measures that assess the Dark Triad in a single inventory: the Dirty Dozen (Jonason & Webster, 2010) and the Short Dark Triad (SD3; Jones & Paulhus, 2014). Whereas recent research has raised concerns about the validity of the Dirty Dozen due to its brevity and lack of discriminant validity at the trait level (e.g., Miller et al., 2012), the SD3 has demonstrated good convergent validity and structural equivalence with the traditional, yet isolated measures (Maples, Lamkin, & Miller, 2014), as well as antisocial outcomes, including sexual coercion and workplace deviance (Jones & Olderbak, 2014; Palmer, Komarraju, Carter, & Karau, 2017). As such, we use the SD3 in this work.

1.2. The dark triad and online behavior

Although the relationship between the Dark Triad traits and online behaviors has received little attention, existent research appears to parallel that in face-to-face settings (Back et al., 2010; Book et al., 2015; Christie & Geis, 1970). We therefore ground our predictions in studies of both offline and online behavior. Crossley, Woodworth, Black, and Hare (2016) examined the Dark Triad in the online domain by investigating the efficacy of negotiation in both face-to-face and computer-mediated interactions. Results indicated that individuals scoring higher in Dark Triad traits were more successful than the average individual in negotiation in face-to-face interactions and also more successful–albeit to a lesser degree– in online interactions.

Cho et al. (2016) conducted a study exploring phishing and the Big Five personality traits (Costa & McCrae, 1992) and end-user susceptibility to phishing emails. They found that high neuroticism was associated with lowered trust and increased perceptions of risk. These findings were especially pronounced when participants were also low in conscientiousness or high in openness to experience. Further, individuals high in agreeableness were more likely to be deceived by phishing emails because they were more trusting and perceived less risk in responding to emails from others. However, overall accuracy did not vary because such trust and risk perceptions were present in response to both benign as well as malicious emails.

If agreeableness is a risk factor for opening malicious emails, then it follows that low agreeableness would be a protective factor for end-users. Because all three Dark Triad traits are low in agreeableness (e.g., Jakobwitz & Egan, 2006; Paulhus & Williams, 2002), it may initially seem that individuals high in any of the Dark Triad traits would not be vulnerable to malicious email attacks. However, narcissism is also associated with overconfidence (Campbell, Goodie, & Foster, 2004) and functional impulsivity (Jones & Paulhus, 2011b). Thus, narcissistic individuals act quickly and often (falsely) believe that they know what they are doing in novel situations. These beliefs create an unrealistic sense of optimism and invulnerability (Farwell & Wohlwend-Lloyd, 1998).

Dark Triad personality research also lends itself to the prediction of adversarial behavior in attackers. Because of their impulsive nature, individuals high in psychopathy are generally poor at tasks involving attention to detail (Newman, 1987). As a consequence, they are less likely to invest time and effort in crafting a single phishing email and more likely to send emails that require little effort and make minimal changes between emails. In contrast, because of their cautious and strategic nature, those high in Machiavellianism tend to plan out their next moves (Czibor & Bereczkei, 2012) and calibrate their strategies based on their audience (Esperger & Bereczkei, 2012).

Based on these findings, we hypothesize that attackers higher in Machiavellianism will engage in more effort to change and adapt emails. In contrast, we expect both psychopathy and narcissism to be linked to less individualized techniques, such as mass standardized “scamming.” Among end-users, we predict that narcissism—but not Machiavellianism or psychopathy—will be linked with higher susceptibility to phishing attacks because of the overconfidence associated with the trait.

2. Methods

The methods described below include only the sections of the original study related to the Dark Triad, attackers' phishing effort and performance, and end-users’ classification of emails. For information about the full study and other measures and findings, please consult Rajivan and Gonzalez (2018).

2.1. Participants

Participant recruitment occurred in two stages. First, 100 participants (49% female) between ages 18 and 75 were recruited from Amazon Mechanical Turk to act as phishing attackers. A second group of 340 participants was later recruited, also from Amazon Mechanical Turk, as end-users (no demographic information was collected in this phase). All participants received $1.50 for participating, and participants in the attacker role had the opportunity to receive up to an additional $4.00 in bonuses for performance. The sample size was determined early in the design by Rajivan and Gonzalez (2018) according to pilot studies to estimate a 95% CI for phishing effort and classification measures.

2.2. Design

The study employed a novel two-phase design. In Phase 1, participants in the attacker role were provided instructions and basic training about phishing and phishing attacks via email. After seeing examples of real phishing emails, participants performed two practice trials of the experimental task, which was to develop phishing emails based on a phishing email template that was randomly assigned to them at the start of the experiment. Attackers were instructed to write phishing emails with two primary goals: (i) to evade spam filter detection and (ii) to persuade end-users to respond. Individuals were told they could evade detection by changing the content of the email (body and subject line). Emails could be edited as much as desired, and no limitations were imposed on the length of each email. Participants were told they could accumulate rewards in each trial based on the success of their phishing email in achieving the two stated goals. After the practice trials, each participant developed eight such phishing emails. At the end of each trial, participants received feedback about their net reward in terms of points. The net reward consisted of both the reward for persuasion and the reward for evasion, but this breakdown was not shown

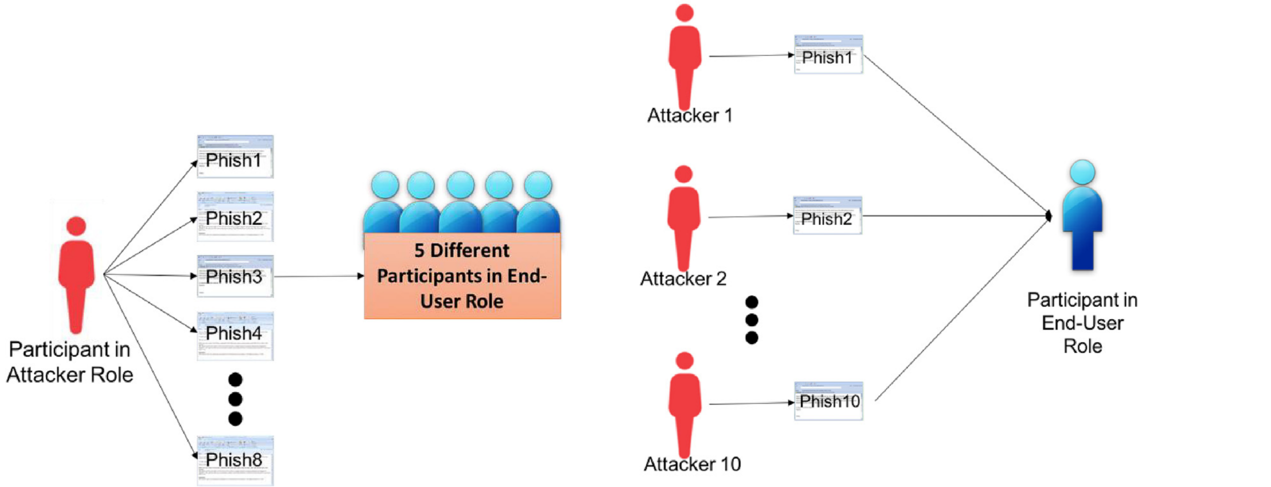
Participant in Attacker Role
Phish1
Phish2
Phish3
Phish4
Phish8
5 Different Participants in End-User Role
Attacker 1
Attacker 2
Attacker 10
Phish1
Phish2
Phish10
Participant in End-User Role

Fig. 1. Visual representation of distribution of phishing Emails to participants in Phase-2 – Figure taken from Rajivan and Gonzalez (2018) (Fig. 3).

to participants.

In Phase 2, participants in the end-user role were presented with a series of 20 emails to evaluate. They were instructed that the goal of the study was to understand how people manage their e-mails. Ten of these emails were benign in nature (i.e., ham emails), while the other 10 were malicious phishing emails created and edited by participants in the attacker role during Phase 1. End-users' stated task was to examine each email with the aim of helping a fictional office manager process her inbox. For each email, they were asked to select one response action they would take to manage it: respond immediately (1); leave the email in the inbox and flag for follow up (2); leave the email in the inbox (3); delete the email (4); delete the email and block the sender (5).

A custom randomization algorithm was used for random assignment of phishing emails generated from phase-1 to participants in end-user role in phase-2. Use of such a randomization algorithm ensured that each end-user participant received 10 unique phishing emails from participants in phase-1. Each phishing email presented to the end-user was created by a different participant in the attacker role. For example, see Fig. 1 the end-user is shown to receive 10 phishing emails, Phish1 to Phish10, from 10 different attackers. Furthermore, each eligible e-mail (with 50-character edits or more to the body of the email) from phase-1 was rated by to five different participants in phase-2. For example, see Fig. 1 where the email identified as "Phish3" from one attacker is shown to be distributed to five different end-users.

Such a conditional random assignment ensured that participants in the end-user role responded to a variety of phishing emails from different participant sources and therefore, less likely to introduce variance from learning effects and other confounds. In total, 635 phishing emails were evaluated.

2.3. Dark triad scores

After completing their respective email tasks, both attackers and end-users were asked to complete the 27-item Short Dark Triad (SD3; Jones & Paulhus, 2014), which assesses psychopathy, Machiavellianism, and narcissism. Each trait was assessed by nine items, and the scores for these nine items were averaged to obtain a composite score for each trait. Each item was scored on a five-point Likert-type scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). Reliability for each Dark Triad construct was acceptable for both the attackers (Machiavellianism $\alpha = 0.85$; narcissism $\alpha = 0.81$; psychopathy $\alpha = .8$) and the end-users (Machiavellianism $\alpha = 0.79$; narcissism $\alpha = 0.80$; psychopathy $\alpha = .78$). Correlations between the traits for attackers (Mach/Narc: $r = 0.53$, Mach/Psyc: $r = 0.55$, Narc/Psyc: $r = 0.58$, all $p < .001$) and end-users (Mach/Narc: $r = 0.35$, Mach/Psyc: $r = 0.53$, Narc/Psyc: $r = 0.36$, all $p < .001$) were similar to those found in previous studies (Jones & Paulhus, 2014).

2.4. Phishing effort

At the attacker level, two behavioral variables assessed phishing effort: number of character edits to the subject line and number of character edits to the body of the email. The number of character edits made in each trial was assessed using the Levenshtein distance, a standard approach to measure the similarity/distance between two given strings (Navarro, 2001). This distance function was used to calculate total number edits based on the number of insertions, deletions, and substitutions made in each trial compared to the email generated by the same participant in the previous trial.

2.5. Phishing classification score

To determine whether the attacker's phishing emails persuaded the end-users to respond, we summed the five different end-user responses collected for each phishing email to create an aggregate email classification score. This aggregate score ranged from 5 (all five end-users chose to immediately respond to the email) to 25 (all five end-users chose to delete the email). For the attacker, a low aggregate phishing classification score meant low loss and high success, and a high phishing classification score meant high loss and low success. The interpretation of the same aggregated classification score was reversed for the end-user. A low phishing classification score meant low performance (high susceptibility), and a high score meant high performance (low susceptibility).

In addition, as each end-user also rated 10 benevolent emails, we took the mean of these 10 classifications to form a single "ham classification" variable. This variable ranged from 1 (an end-user chose to immediately respond to all 10 ham emails) to 5 (an end-user chose to delete all 10 ham emails). The calculation of this variable is important to determine the ability of our end-user participants to discriminate between benevolent and malevolent emails.

3. Results

Table 1 presents the average Dark Triad scores of both, attackers and end-users, as well as end-users' average classification scores for each type of email: ham and phishing. These variables had approximately normal distributions and replicate descriptive findings from previous studies (see Jones & Paulhus, 2014). Attackers' phishing effort and loss were used as outcome measures to model the relationship

Table 1
Descriptive statistics for attackers and end users.

	M	*SD*	Skew	Kurtosis
Psychopathy				
End-User	2.22	0.66	0.05	−0.61
Attacker	2.15	0.76	0.87	0.88
Machiavellianism				
End-User	3.10	0.65	0.01	−0.25
Attacker	3.15	0.74	−0.28	1.04
Narcissism				
End-User	2.71	0.70	0.19	0.11
Attacker	2.63	0.72	0.23	−0.71
End-User: Ham Email Ratings	2.70	0.54	0.37	0.85
End-User: Phishing Email Ratings	3.11	0.75	0.01	−0.27
Attacker: Performance	15.37	3.10	−0.01	0.10
Attacker: Effort on Body Edits	253.36	227.71	5.85	59.80
Attacker: Effort on Subject Edits	15.80	14.89	0.80	0.60

Note. Dark Triad traits were measured on a scale of 1–5, with 5 indicating the highest degree of that trait. The possible range for end-user email ratings and attacker performance was 5–25. Attacker effort was assessed by the Levenshtein distance between the number of edits made between trials.

between individual Dark Triad traits and phishing behaviors.

3.1. Attacker's dark triad scores and phishing effort

The number of body edits ranged from 50 to 2885 characters, and the distribution of edits was overdispersed (see Table 1). Although Poisson regressions are generally conducted for the analysis of count data, participants were highly variable in the number of character edits they made to both the bodies and subjects of their emails. The overdispersion of the variables required the use of negative binomial regressions (Gardner, Mulvey, & Shaw, 1995) to identify the effect of Dark Triad traits on phishing effort as defined by the amount of body edits made. A negative binomial regression is a generalized linear model specific for modeling overdispersed count variables by incorporating an additional parameter, the dispersion parameter that allows for the variance to differ from the mean, thus relaxing the assumptions of a Poisson distribution for count variables (Gardner et al., 1995).

Fig. 2 shows the relationships between attackers' psychopathy, narcissism, and Machiavellianism scores and the number of edits made to the body of the email. These relationships were negative for psychopathy and narcissism but positive for Machiavellianism. Our results suggest that attackers higher in Machiavellianism made more edits to the body of the email, $\beta = 0.16$, $SE = 0.03$, $Est/SE = 5.72$, $p < .001$. Both narcissism and psychopathy were significantly negatively related to body email edits, Narcissism: $\beta = -0.10$, $SE = 0.02$, $Est/SE = -4.33$, $p < .001$; Psychopathy: $\beta = -0.09$, $SE = 0.03$, $Est/SE = -3.49$, $p < .001$. Specifically, the beta weights provided are standardized slope estimations of the change in the number of body edits for every one unit increase in the given Dark Triad Trait. Fig. 2 provides a visual representation of these slopes.

Fig. 3 shows the relationships between the attackers' Dark Triad traits and the number of edits made to the subject line of the email. The fitted line in the graph plots the likelihood of making zero subject edits. Because 32.4% of emails had zero subject edits made, subject line edits were analyzed using a zero-inflated negative binomial regression. To account for the abundance of zeros in the model, a zero-inflated negative binomial regression conducts two analyses: a general linear regression model with over-abundant zeros excluded from the model, and a binary logistic regression predicting the likelihood of making zero edits compared to making at least one edit (Zuur, Ieno, Walker, Saveliev, & Smith, 2009).

This analysis indicated that attackers higher in psychopathy were less likely to make any edits to the subject of the email, $\beta = 0.30$, $SE = 0.06$, $Est/SE = 4.77$, $p < .001$. Note that the positive beta comes from the logistic regression (conducted as part of zero-inflated negative binomial regression) and implies that attackers higher in psychopathy were more likely to make zero edits to the subject line. In contrast, individuals higher in narcissism were more likely to make at least one edit to the subject of the email, $\beta = -0.15$, $SE = 0.07$, $Est/SE = -2.22$, $p = .03$. There was no evidence to suggest that Machiavellianism was related to subject line edits, $\beta = -0.01$, $SE = 0.07$, $Est/SE = -0.10$, $p = .92$.

After accounting for the overabundance of zero subject edits, the zero-inflated negative binomial regression suggested no significant relationship between any of the attacker Dark Triad scores and making additional edits to the subject line, Psychopathy: $\beta = -0.02$; Narcissism: $\beta = -0.03$; Machiavellianism: $\beta = 0.01$; all p-values $>$ .05. Thus, these personality differences predicted the binary decision to either make or not make a change to the subject line, but they did not predict additional edits after participants made the initial decision to edit the subject line.

3.2. Attacker's dark triad and phishing classification score

Analyses predicting the phishing classification score were conducted in a three-level multilevel framework using MPlus 7.1 (Muthen & Muthen, 2012) with the end-user as the individual unit of analysis, the edits made to the phishing emails as the level 2 clustering variable, and attackers' Dark Triad scores as the level 3 clustering variable. See Fig. 4 for a visualization of the data structure and variables at each level. A null model indicated intra-class correlations of 0.128 and 0.118, indicating that 12.8% of the variance in individual classifications of emails was due to variations within the emails themselves and that 11.8% was due to variability at the attacker level. Note that these intraclass correlations indicate all of the between and within level variability in a null model, or a model without any predictors. Thus, the percentages of variance indicated above could be due to the variables we estimate in our models or additional variables that we did not anticipate or account for.

A random intercepts three-level model was conducted using a Bayesian estimator. The estimation method used Gibbs sampling and two Markov chain Monte Carlo chains. This estimation method is recommended for three-level models since likelihood-based methods have been found to bias point estimates when three levels are analyzed (Browne & Draper, 2006). The use of a single three-level model allowed us to parse out the variance attributable to each predictor variable across levels and account for all unit clustering. Although separate analyses would have been viable to differentially assess attacker loss and end-user susceptibility, results from these models would potentially have been confounded by the excluded levels. At level 1, predictor variables were the end-user Dark Triad scores and their classification of ham emails. At level 2, predictor variables were the number of edits made to the subject and body of the email. At level 3, the predictor variables were the attacker scores on the Dark Triad constructs.

Table 2 shows the overall findings for the three-level model. At the attacker level, no variables were significantly predictive of phishing classification score. Attacker narcissism, psychopathy, and Machiavellianism were not significantly related to the classification score, all p-values $>$ .05. Further, the number of character edits to both the body and subject lines of the phishing emails was not directly predictive of email classification, all p-values $>$ .05. Therefore, even though personality differences in attackers did predict effort in terms of edits to the body of the phishing email, neither these personality differences nor these character edits significantly influenced how end-users classified the emails.

3.3. End-users' dark triad and phishing classification score

To assess whether specific personality characteristics of end-users predicted classification score, we examined the same three-level model

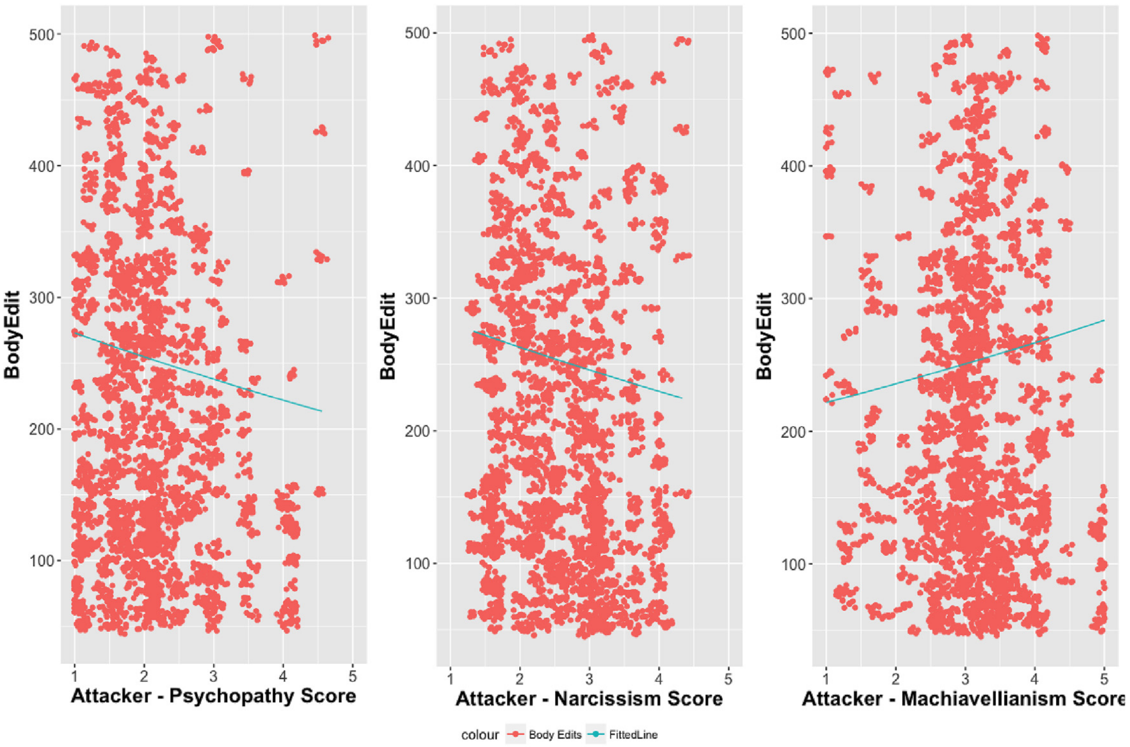

Fig. 2. Data visualization of the relationship between the Dark Triad and edits to the body of the Phishing emails.

discussed previously (results are located in Table 2). Although these analyses could have been examined with a single-level multiple regression, the use of a single three-level model allowed us to simultaneously isolate the effects at each level of analysis.

Both end-user narcissism and end-user psychopathy were significantly predictive of email classification. Specifically, both were associated with greater susceptibility to phishing emails, Narcissism: $\beta = -0.18$, $SE = 0.04$, 95% CI: [-0.25, -0.11]; Psychopathy: $\beta = -0.11$, $SE = 0.04$, 95% CI: [-0.19, -0.02]. Exact p-values are not reported, as these analyses were carried out using a Bayesian estimator. These findings indicate that for every one-unit increase in narcissism/psychopathy scores, end-users rated an email an average of .18/0.11 points less suspicious (i.e., an increased likelihood to respond to the email). Further, individuals who more often classified the ham emails

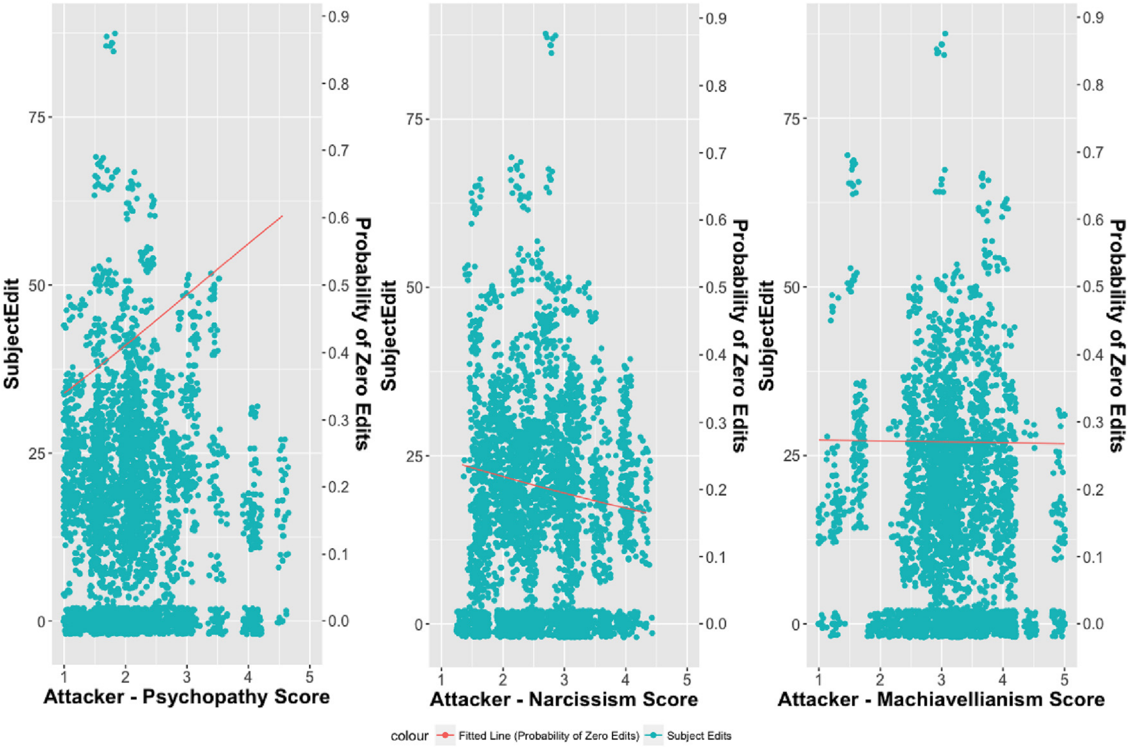

Fig. 3. Data visualization of the relationship between Dark Triad scores and the likelihood of making zero edits to the subject line of the Phishing emails.

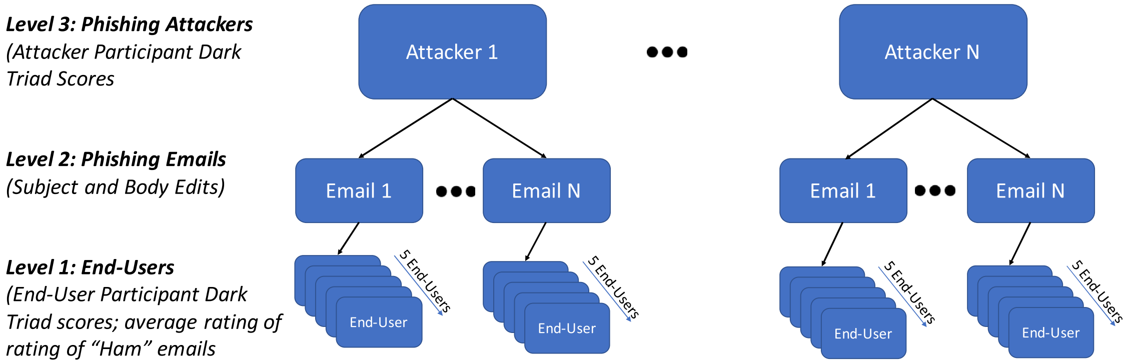
Level 3: Phishing Attackers
(Attacker Participant Dark Triad Scores
Level 2: Phishing Emails
(Subject and Body Edits)
Level 1: End-Users
(End-User Participant Dark Triad scores; average rating of rating of "Ham" emails
Attacker 1
Attacker N
Email 1
Email N
Email 1
Email N
End-User
5 End-Users

Fig. 4. Visualization of a three-level data structure.

Table 2
Three-level random intercepts model results.

Parameter	*β*	*SE*	95% CI
Fixed Effects - Level 1			
Intercept	1.76*	0.33	[1.12, 2.42]
End-user Narcissism	−0.18*	0.04	[-0.25, −0.11]
End-user Psychopathy	−0.11*	0.04	[-0.19, −0.02]
End-user Machiavellianism	−0.01	0.04	[-0.09, 0.08]
End-user Ham Email Rating	0.75*	0.04	[0.66, 0.83]
Fixed Effects - Level 2			
Subject Edits	0.00	0.00	[0.00, 0.01]
Body Edits	0.00	0.00	[0.00, 0.00]
Fixed Effects - Level 3			
Attacker Narcissism	0.04	0.11	[-0.17, 0.26]
Attacker Psychopathy	−0.02	0.10	[-0.22, 0.18]
Attacker Machiavellianism	−0.02	0.10	[-0.23, 0.18]
Variance Components			
Residual Variance Level 1	1.46*	0.04	[1.38, 1.54]
Residual Variance Level 2	0.26*	0.03	[0.20, 0.33]
Residual Variance Level 3	0.27*	0.06	[0.18, 0.40]

*$p < .05$.

as suspicious were also more likely to classify the phishing emails as suspicious, $\beta = 0.75$, $SE = 0.04$, 95% CI: [0.66, 0.83]. Specifically, for every one-unit increase in "ham classification", end-users rated phishing emails an average of 0.75 points more suspicious (i.e. a lower likelihood of response to the email). This relationship suggests that, overall, end-users were not very successful at discriminating between ham and phishing emails. Instead, the majority of their email classifications were based on thresholds of skepticism and suspiciousness rather than an ability to distinguish a phishing email from a ham email.

3.4. Dyadic interactions between attacker and end-user personality characteristics

Although the three-level model did not suggest that attacker personality characteristics contributed to phishing performance, we were still interested in the possibility that personality characteristics might interact between attackers and end-users. A cross-level interaction effect assesses whether level 2 grouping factors (e.g., attacker personality characteristics) can explain variance across group slopes through moderating the relationship between level 1 variables across level 2 clusters (e.g., different attackers; Aguinis, Gottfredson, & Culpepper, 2013). Due to the complexity of assessing slopes as free-to-vary random effects compared to fixed effects, these relationships cannot be measured in a full three level model but were instead tested post hoc in a separate two-level model using a maximum likelihood estimation method. This test of cross level interactions examined if and how attacker personality characteristics moderated the relationship between end-user personality characteristics and phishing email classification. In order to test cross level interactions, the multilevel model must assume that the variance of slopes across groups is different from zero. In the context of our study, this assumption means that the relationship between end-user personality characteristics and phishing email classifications changes across different attackers.

A two-level model was built in which the attacker was the level 2 clustering variable and the end-users were the individual unit of analysis. Results from this model can be found in Table 3. Only one cross-level interaction approached significance, suggesting that attacker narcissism marginally moderated the relationship between end-user narcissism and phishing email classification, $\beta = 0.10$, $SE = 0.05$, $p = .07$. The positive estimate indicates that for every one-unit increase in attacker narcissism scores, the slope of the relationship between end-user narcissism and phishing classifications increases by 0.10. Thus, end-user narcissism is marginally more strongly related to the email classification score as the attacker's narcissism score increases. Because the slope of the relationship between end-user narcissism and phishing classifications was negative (see Table 2), this finding suggests that as attacker narcissism scores increase, end-users high in narcissism become marginally more susceptible to phishing emails, resulting in lower end-user performance.

To further explore the dyadic relationship between attacker and end-user dark triad personalities, we created a personality delta (Δ) variable, which measures the difference in Dark Triad scores between attackers and end-users across all three constructs. A logistic regression was conducted to plot the relationship between this personality delta

Table 3
Two-level random intercepts random slopes model results: Attackers and end-users.

Parameter	*β*	*SE*	Two-Tailed *p*-value
Fixed Effects - Level 2			
Attacker Psychopathy	−0.03	0.10	.72
Attacker Narcissism	0.04	0.11	.69
Attacker Machiavellianism	−0.02	0.09	.87
Cross-Level Interactions: End-User Narcissism			
Attacker Psychopathy	0.05	0.06	.40
Attacker Narcissism	0.10	0.05	.07
Attacker Machiavellianism	−0.03	0.05	.47
Cross-Level Interactions: End-User Psychopathy			
Attacker Psychopathy	0.06	0.08	.43
Attacker Narcissism	−0.01	0.07	.86
Attacker Machiavellianism	−0.06	0.08	.48
Cross-Level Interactions: End-User Machiavellianism			
Attacker Psychopathy	−0.10	0.07	.18
Attacker Narcissism	−0.05	0.07	.46
Attacker Machiavellianism	−0.00	0.06	.95
Variance Components			
Within-Level Variance	1.64	0.07	< .001
Intercept Variance	0.28	0.05	< .001
Slope – End-User Narcissism	0.00	0.03	.94
Slope – End-User Psychopathy	0.04	0.02	.12
Slope – End-User Machiavellianism	0.02	0.02	.40

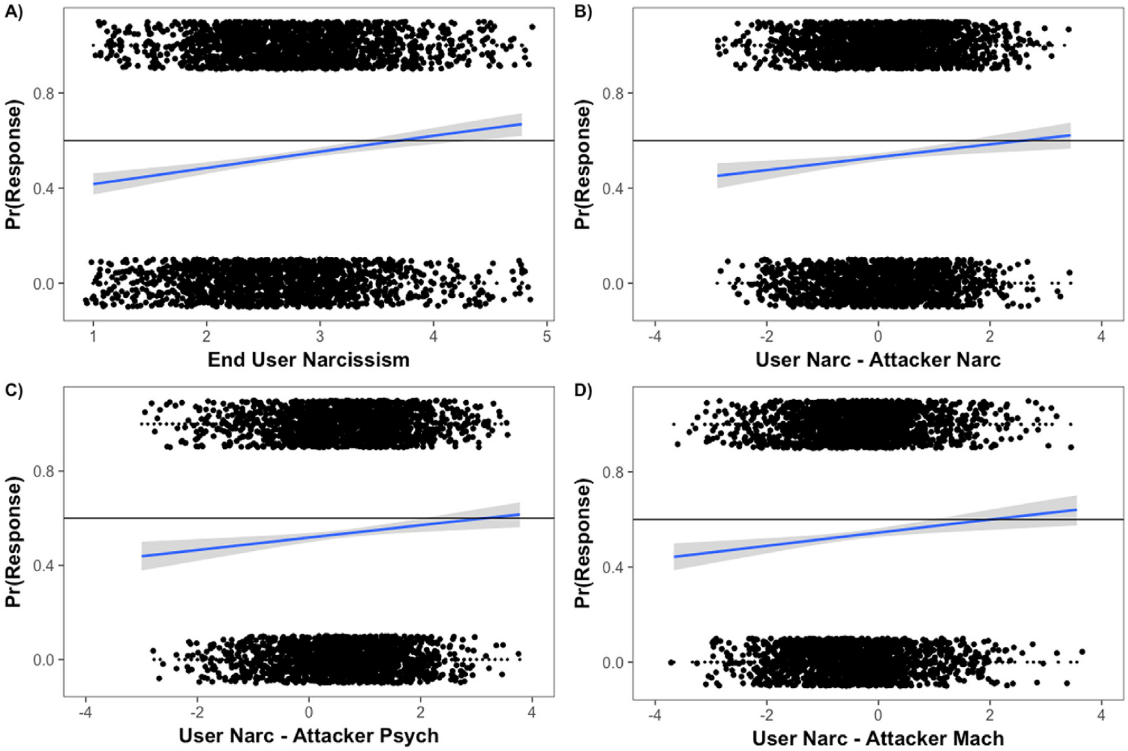

Fig. 5. Dichotomized visualization of the likelihood to respond to a phishing email based on end-user narcissism and differences between end-user and attacker Dark Triad traits. Classification scores were artificially separated into "respond" or "not respond" (dots correspond to each classification), and a slope of the probability of response was estimated using the difference scores presented in each subfigure. No statistical inferences have been made from this visualization represented in this figure.

measure and the likelihood of end-user response to phishing emails. Fig. 5 shows the plots specifically for end-user narcissism. Plots for other end-user constructs were neglected due to their lack of relationship to phishing classification. In Fig. 5, the x-axis in plots b, c and d denotes the difference between the end-user scores for narcissism and the attacker scores for the three dark triad constructs, calculated by subtracting the latter from the former. For example, the x-axis in Fig. 5c denotes the ΔNP: the difference between end-user narcissism and attacker psychopathy. Instances with high ΔNP indicate that an end-user who scored high on narcissism responded to a phishing email from an attacker who scored low on psychopathy. In contrast, instances with low ΔNP indicate that an end-user who scored low on narcissism responded to phishing email from an attacker who scored high on psychopathy. Similar interpretations apply to other combinations of ΔNN (end-user narcissism vs. Attacker narcissism) and ΔNM (end-user narcissism vs. Attacker Machiavellianism). For comparison, Fig. 5a plots the direct relationship between end-user narcissism and the likelihood of responding to a phishing email.

As can be seen in Fig. 5, regardless of the attacker scores on the different Dark Triad constructs, the higher the end-user was in narcissism (or as the difference between the two became more positive), the more likely they were to respond to phishing emails. This relationship occurred at about the same likelihood as when attacker Dark Triad scores were not included (refer to Fig. 5a). Further, Fig. 5b shows a slightly shallower slope than both Fig. 5c and d. This finding lends visual support to the results from the cross-level interaction in the multilevel model: as attacker narcissism increases, individuals higher in end-user narcissism become marginally more susceptible to those specific phishing emails, resulting in a shallower slope for the difference scores.

4. Discussion

Overall, our findings demonstrate that although attackers' scores on the Dark Triad do not predict the effectiveness of their phishing emails, end-users' Dark Triad scores do. Furthermore, our results suggest a marginally significant relationship of dyadic reciprocity between narcissistic end-users who are more susceptible to phishing emails and narcissistic attackers. These results expand initial findings regarding online manipulation and the Dark Triad (Crossley et al., 2016) in the following ways.

First, our results revealed novel relationships between attackers' behavior and personalities. We found that attacker levels of Machiavellianism were linked to how much effort they put into writing their phishing emails, measured by the number of changes they made in the body of the email. This finding reflects the caution and planning that has been associated with this trait (Jones & Paulhus, 2011a). Narcissism, on the other hand, was associated with fewer changes to the body of the phishing email, and psychopathy was negatively correlated with the number of changes to the subject line of the phishing email. These findings are consistent with past research that demonstrates the unique relationships narcissism and psychopathy have with different forms of impulsivity (Jones & Paulhus, 2011b). Individuals high in psychopathy may simply be in a hurry to email as many end-users as possible, hoping that some individuals will fall victim. The association between narcissism and fewer changes to the body of the phishing emails may be explained by their overconfidence. Specifically, such individuals may believe that their superior skills necessitate little changes to evade detection.

Second, our results also revealed novel relationships between end-user behavior and narcissism. Among end-users, narcissism was associated with greater vulnerability to phishing emails. There are several potential explanations for this finding. The first is associated with the overconfidence inherent to narcissism (Campbell et al., 2004). Such individuals are unrealistically optimistic about potential outcomes and their ability to deal with tricky situations, which is driven by their superior sense of self (Farwell & Wohlwend-Lloyd, 1998). Such a superior sense of self may also translate into blaming others and externalizing fault, which may interfere with learning and awareness of future attacks (Dutt, Ahn, & Gonzalez, 2013). Because these individuals believe that they are more knowledgeable than they actually are across a wide variety of contexts (Paulhus & Williams, 2002), they may (falsely) believe that they would detect a phishing attempt if they saw one. Second, individuals high in narcissism are functionally impulsive (Jones & Paulhus, 2011b). As a consequence, they may see a potentially interesting or rewarding email and pursue it without much thought to the potential consequences.

Third, our results revealed a marginally significant dyadic relationship between narcissistic individuals. Narcissistic end-users were more susceptible to phishing emails that originated from narcissistic attackers. It is important to note that end-users and attackers never directly interacted in this experiment and it is therefore in the interests of future studies to research *direct* dyadic interactions among the different personalities in adversarial settings. Furthermore, participants may have been hyper-aware of potential deception because of the nature of the study. Thus, realistic and naturalistic observation are needed to bolster the present findings. Future research should also explore "spear-phishing" attacks, which include information that can inform a targeted attack. Specific dyadic relationships between the personalities of end-users and attackers may be important to explaining spear-phishing success. Nevertheless, the present research represents a good first step into understanding both the attack patterns and vulnerabilities that may be present in individuals who are high in different Dark Triad traits. This understanding may lead organizations to best tailor specific interventions towards overconfident end-users, and provide spam filters with increased information about who might be most vulnerable to specific types of phishing attempts.

This study is not without its limitations. As mentioned above, due to the nature of this study, end-user participants may have been hyper-aware of potential deception, and thus more vigilant in their ratings of each email than they would be in their natural work environment. Therefore, future research should aim to enhance ecological validity. Further, the emails that end-users were evaluating were on behalf of another person (see Rajivan & Gonzalez, 2018) rather than emails directed at the participants. Decision making strategies and phishing vulnerabilities may shift dependent on if the decision is being made for the self vs. Another. This question should also be addressed in future work. In addition, although this study investigated the end-users' evaluations of each email, we did not include metrics to identify how these participants were making their decisions. Therefore, it remains unclear as to what aspects of the emails participants were flagging as suspicious across both phishing and ham emails.

In sum, this work identifies some individual differences associated with attacker and end-user behavior that may help identify potential attackers and understand who is at greatest risk for falling prey to an attack in an organization. Namely, Dark Triad traits can predict how much effort an individual would put into designing a phishing email, and individuals who are narcissistic may be especially at risk of compromising not only personal security, but also an organization's security. Although automated email filtering programs aim to eliminate the presence of spam and phishing emails presented to individuals, malicious attacks still manage to breach these filters. Understanding the types of people who create these emails, how they create them, and which individuals may be more vulnerable to falling for these attacks is an essential step to limit security breaches. This knowledge can have real world implications in multiple areas of business and security. One such implication is the potential implementation of training programs to better differentiate between phishing and ham emails, particularly for those individuals high in narcissism. Further, current laws and regulations on data protection and privacy may be better informed by the findings presented in this study. Phishing emails present a complex form of data breach within corporate and governmental organizations, as the cause of such a breach would be due to human susceptibility and error, rather than flaws within computerized safe guards. Therefore, a clear understanding of the patterns utilized by attackers and vulnerabilities presented by end-users is essential to minimizing the risks that phishing presents to individuals, businesses, and government.

Acknowledgements

This research was supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. The authors thank Nalyn Sriwattanakomen for editing the manuscript.

References

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management, 39*(6), 1490–1528. http://dx.doi.org/10.1177/0149206313478188.

Anderson, W. L. (2010). Cyber stalking (cyberbullying)-proof and punishment. *Insights to a Changing World Journal, 4*(3), 18–23.

Azizli, N., Atkinson, B. E., Baughman, H. M., Chin, K., Vernon, P. A., Harris, E., et al. (2016). Lies and crimes: Dark Triad, misconduct, and high-stakes deception. *Personality and Individual Differences, 89*, 34–39. http://dx.doi.org/10.1016/j.paid.2015.09.034.

Back, M. D., Schmukle, S. C., & Egloff, B. (2010). Why are narcissists so charming at first sight? Decoding the narcissism–popularity link at zero acquaintance. *Journal of Personality and Social Psychology, 98*(1), 132–145. http://dx.doi.org/10.1037/a0016338.

Bereczkei, T. (2015). The manipulative skill: Cognitive devices and their neural correlates underlying Machiavellian's decision making. *Brain and Cognition, 99*, 24–31. http://dx.doi.org/10.1016/j.bandc.2015.06.007.

Book, A., Methot, T., Gauthier, N., Hosker-Field, A., Forth, A., Quinsey, V., et al. (2015). The mask of sanity revisited: Psychopathic traits and affective mimicry. *Evolutionary Psychological Science, 1*(2), 91–102. http://dx.doi.org/10.1007/s40806-015-0012-x.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian analysis, 1*(3), 473–514. http://dx.doi.org/10.1214/06-BA117.

Campbell, W. K., Goodie, A. S., & Foster, J. D. (2004). Narcissism, confidence, and risk attitude. *Journal of Behavioral Decision Making, 17*(4), 297–311. http://dx.doi.org/10.1002/bdm.475.

Cho, J. H., Cam, H., & Oltramari, A. (2016). Effect of personality traits on trust and risk to phishing vulnerability: Modeling and analysis. March *2016 IEEE International multidisciplinary conference on cognitive methods in situation awareness and decision support (CogSIMA)* (pp. 7–13). IEEE. http://dx.doi.org/10.1109/COGSIMA.2016.7497779.

Christie, R., & Geis, F. (1970). *Studies in machiavellianism.* New York, NY: Academic Press.

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13*(6), 653–665. http://dx.doi.org/10.1016/0191-8869(92)90236-I.

Crossley, L., Woodworth, M., Black, P. J., & Hare, R. (2016). The dark side of negotiation: Examining the outcomes of face-to-face and computer-mediated negotiations among dark personalities. *Personality and Individual Differences, 91*, 47–51. http://dx.doi.org/10.1016/j.paid.2015.11.052.

Czibor, A., & Bereczkei, T. (2012). Machiavellian people's success results from monitoring their partners. *Personality and Individual Differences, 53*(3), 202–206. http://dx.doi.org/10.1016/j.paid.2012.03.005.

Dutt, V., Ahn, Y.-S., & Gonzalez, C. (2013). Cyber situation awareness: Modeling detection of cyber-attacks with instance-based learning theory. *Human Factors, 55*(3), 605–618. http://dx.doi.org/10.1177/0018720812464045.

Emmons, R. A. (1987). Narcissism: Theory and measurement. *Journal of Personality and Social Psychology, 52*(1), 11–17. http://dx.doi.org/10.1037/0022-3514.52.1.11.

Esperger, Z., & Bereczkei, T. (2012). Machiavellianism and spontaneous mentalization: One step ahead of others. *European Journal of Personality, 26*(6), 580–587. http://dx.doi.org/10.1002/per.859.

Farwell, L., & Wohlwend-Lloyd, R. (1998). Narcissistic processes: Optimistic expectations, favorable self-evaluations, and self-enhancing attributions. *Journal of Personality, 66*(1), 65–83. http://dx.doi.org/10.1111/1467-6494.00003.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin, 118*(3), 392–404. http://dx.doi.org/10.1037/0033-2909.118.3.392.

Hong, J. (2012). The state of phishing attacks. *Communications of the ACM, 55*(1), 74–81. http://dx.doi.org/10.1145/2063176.2063197.

Im, G. P., & Baskerville, R. L. (2005). A longitudinal study of information system threat categories: The enduring problem of human error. *ACM SIGMIS - Data Base: The DATABASE for Advances in Information Systems, 36*(4), 68–79. http://dx.doi.org/10.1145/1104004.1104010.

Jakobwitz, S., & Egan, V. (2006). The Dark Triad and normal personality traits. *Personality and Individual Differences, 40*(2), 331–339. http://dx.doi.org/10.1016/j.paid.2005.07.006.

Jauk, E., Neubauer, A. C., Mairunteregger, T., Pemp, S., Sieber, K. P., & Rauthmann, J. F. (2016). How alluring are dark personalities? The Dark Triad and attractiveness in speed dating. *European Journal of Personality, 30*(2), 125–138. http://dx.doi.org/10.1002/per.2040.

Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: A concise measure of the dark triad. *Psychological Assessment, 22*(2), 420–432. http://dx.doi.org/10.1037/a0019265.

Jonason, P. K., & Webster, G. D. (2012). A protean approach to social influence: Dark Triad personalities and social influence tactics. *Personality and Individual Differences, 52*(4), 521–526. http://dx.doi.org/10.1016/j.paid.2011.11.023.

Jones, D. N., & Olderbak, S. G. (2014). The associations among dark personalities and sexual tactics across different scenarios. *Journal of Interpersonal Violence, 29*(6), 1050–1070. http://dx.doi.org/10.1177/0886260513506053.

Jones, D. N., & Paulhus, D. L. (2011a). Differentiating the dark triad within the interpersonal circumplex. In L. M. Horowitz, & S. N. Strack (Eds.). *Handbook of interpersonal theory and research* (pp. 249–267). New York, NY: Guilford.

Jones, D. N., & Paulhus, D. L. (2011b). The role of impulsivity in the Dark Triad of personality. *Personality and Individual Differences, 51*(5), 679–682. http://dx.doi.org/10.1016/j.paid.2011.04.011.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3) a brief measure of dark personality traits. *Assessment, 21*(1), 28–41. http://dx.doi.org/10.1177/1073191113514105.

Maples, J. L., Lamkin, J., & Miller, J. D. (2014). A test of two brief measures of the dark triad: The dirty dozen and short dark triad. *Psychological Assessment, 26*(1), 326–331. http://dx.doi.org/10.1037/a0035084.

McHoskey, J. (1995). Narcissism and machiavellianism. *Psychological Reports, 77*(3), 755–759. http://dx.doi.org/10.2466/pr0.1995.77.3.755.

McHoskey, J. W., Worzel, W., & Szyarto, C. (1998). Machiavellianism and psychopathy. *Journal of Personality and Social Psychology, 74*(1), 192–210. http://dx.doi.org/10.1037/0022-3514.74.1.192.

Miller, J. D., Few, L. R., Seibert, L. A., Watts, A., Zeichner, A., & Lynam, D. R. (2012). An examination of the dirty dozen measure of psychopathy: A cautionary tale about the costs of brief measures. *Psychological Assessment, 24*(4), 1048–1052. http://dx.doi.org/10.1037/a0028583.

Muthen, L. K., & Muthen, B. O. (2012). *Mplus statistical modeling Software: Release 7.1 [computer software].* (Los Angeles, CA).

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys, 33*(1), 31–88. http://dx.doi.org/10.1145/375360.375365.

Newman, J. P. (1987). Reaction to punishment in extraverts and psychopaths: Implications for the impulsive behavior of disinhibited individuals. *Journal of Research in Personality, 21*(4), 464–480. http://dx.doi.org/10.1016/0092-6566(87)90033-X.

Palmer, J. C., Komarraju, M., Carter, M. Z., & Karau, S. J. (2017). Angel on one shoulder: Can perceived organizational support moderate the relationship between the Dark Triad traits and counterproductive work behavior? *Personality and Individual Differences, 110*, 31–37.

Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality, 36*(6), 556–563. http://dx.doi.org/10.1016/S0092-6566(02)00505-6.

Rajivan, P., & Gonzalez, C. (2018). Creative persuasion: A study on adversarial behaviors and strategies in phishing attacks. *Frontiers in Psychology, 9*, 135. http://dx.doi.org/10.3389/fpsyg.2018.00135.

Stanton, J. M., Stam, K. R., Mastrangelo, P., & Jolton, J. (2005). Analysis of end user security behaviors. *Computers & Security, 24*(2), 124–133. http://dx.doi.org/10.1016/j.cose.2004.07.001.

Williams, K. M., Paulhus, D. L., & Hare, R. D. (2007). Capturing the four-factor structure of psychopathy in college students via self-report. *Journal of Personality Assessment, 88*(2), 205–219. http://dx.doi.org/10.1080/00223890701268074.

Willison, R., & Backhouse, J. (2006). Opportunities for computer crime: Considering systems risk from a criminological perspective. *European Journal of Information Systems, 15*(4), 403–414. http://dx.doi.org/10.1057/palgrave.ejis.3000592.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). Zero-truncated and zero-inflated models for count data. *Mixed effects models and extensions in ecology with R* (pp. 261–293). New York, NY: Springer.