http://pubsonline.informs.org/journal/isre

INFORMATION SYSTEMS RESEARCH
Vol. 32, No. 2, June 2021, pp. 410–436
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

The Phishing Funnel Model: A Design Artifact to Predict User Susceptibility to Phishing Websites

Ahmed Abbasi,[a] David Dobolyi,[a] Anthony Vance,[b] Fatemeh Mariam Zahedi[c]

[a] Mendoza College of Business, University of Notre Dame, Notre Dame, Indiana 46556; [b] Fox School of Business, Temple University, Philadelphia, Pennsylvania 19122; [c] Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53202

Contact: aabbasi@nd.edu, https://orcid.org/0000-0001-7698-7794 (AA); ddobolyi@nd.edu, https://orcid.org/0000-0002-9493-3447 (DD); anthony@vance.name, https://orcid.org/0000-0002-4554-6176 (AV); zahedi@uwm.edu (FMZ)

Received: August 14, 2017
Revised: November 14, 2018; February 16, 2020; June 16, 2020
Accepted: July 12, 2020
Published Online in Articles in Advance: February 15, 2021

https://doi.org/10.1287/isre.2020.0973

Abstract. Phishing is a significant security concern for organizations, threatening employees and members of the public. Phishing threats against employees can lead to severe security incidents, whereas those against the public can undermine trust, satisfaction, and brand equity. At the root of the problem is the inability of Internet users to identify phishing attacks even when using anti-phishing tools. We propose the phishing funnel model (PFM), a design artifact for predicting user susceptibility to phishing websites. PFM incorporates user, threat, and tool-related factors to predict actions during four key stages of the phishing process: *visit, browse, consider legitimate,* and *intention to transact*. We used a support vector ordinal regression with a custom kernel encompassing a cumulative-link mixed model for representing users' decisions across funnel stages. We evaluated the efficacy of PFM in a 12-month longitudinal field experiment in two organizations involving 1,278 employees and 49,373 phishing interactions. PFM significantly outperformed competing models/methods by 8%–52% in area under the curve, correctly predicting visits to high-severity threats 96% of the time—a result 10% higher than the nearest competitor. A follow-up three-month field study revealed that employees using PFM were significantly less likely to interact with phishing threats relative to comparison models and baseline warnings. Furthermore, a cost-benefit analysis showed that interventions guided by PFM resulted in phishing-related cost reductions of nearly $1,900 per employee more than comparison prediction methods. These results indicate strong external validity for PFM. Our findings have important implications for practice by demonstrating (1) the effectiveness of predicting user susceptibility to phishing as a real-time protection strategy, (2) the value of modeling each stage of the phishing process together, rather than focusing on a single user action, and (3) the considerable impact of anti-phishing tool and threat-related factors on susceptibility to phishing.

History: Olivia Sheng, Senior Editor; Sam Ransbotham, Associate Editor.

Funding: This work was funded by the Division of Information and Intelligent Systems of the National Science Foundation [Grant IIS-1816005], by the Division of Computing and Communication Foundations of the National Science Foundation [Grant CCF-1629450], and by the Division of Computer and Network Systems [Grant CNS-1049497] and the Division of Advanced Cyberinfrastructure [Grant ACI-1443019].
Supplemental Material: The online appendices are available at https://doi.org/10.1287/isre.2020.0973.

Keywords: phishing susceptibility • design science • predictive analytics • online security • longitudinal field experiment

1. Introduction

Phishing—a type of semantic attack that exploits human as opposed to software vulnerabilities (Schneier 2000, Hong 2012)—is one of the most prevalent forms of cybercrime, impacting more than 40 million Internet users every year (Symantec 2012, McAfee 2013, Verizon 2016). Phishing consistently ranks as one of the top security concerns facing IT managers not only because of the number of employees falling prey to phishing attacks within organizations (Bishop et al. 2009, Siponen and Vance 2010, Gartner 2011, Cummings et al. 2012) but also because brand equity and trust are tarnished when customers are targeted by spoof (i.e., fraudulent replica) websites (Hong 2012). The average 10,000-employee company spends approximately $3.7 million annually combating phishing attacks (Korolov 2015).

Several studies have highlighted the markedly poor performance of Internet users when asked to differentiate legitimate websites from phishing or avoid transacting with phishing websites (Grazioli and Jarvenpaa 2000, Jagatic et al. 2007, Li et al. 2014). Prior work has

shown that users are unable to correctly identify phishing websites between 40% and 80% of the time (Grazioli and Jarvenpaa 2000, Dhamija et al. 2006, Herzberg and Jbara 2008, Abbasi et al. 2012a) and that more than 70% of users are willing to transact with phishing websites (Grazioli and Jarvenpaa 2000, Jagatic et al. 2007).

One potential solution to this problem is the use of anti-phishing tools including web browser security toolbars and proprietary toolbars and plug-ins (Li and Helenius 2007, Abbasi et al. 2010, Zhang et al. 2014). Even when using these tools, however, phishing success rates remain high because users often explain away or disregard tool warnings (Wu et al. 2006, Sunshine et al. 2009, Jensen et al. 2010, Abbasi et al. 2012a, Akhawe and Felt 2013). One reason for this failure may be that users do not perceive anti-phishing tool warning as personalized to themselves (Chen et al. 2011).

This study takes a different approach from past anti-phishing tools in that rather than predicting whether a link or website is a phishing attack, we seek to accurately predict users' phishing susceptibility (Downs et al. 2006, Bravo-Lillo et al. 2011). We define phishing susceptibility as *the extent to which a user interacts with a phishing attack*. Such a solution would (1) promote better use of security technologies by addressing factors contributing to user-tool dissonance via personalized real-time warnings, (2) provide personalized access controls and data security policies that reflect users' predicted susceptibility levels, and (3) adapt to changes in high-susceptibility factors that occur over time.

Accordingly, the research objective of this study is *to develop a design artifact for predicting user susceptibility to phishing websites*. We adopted the design science paradigm (Hevner et al. 2004) to guide the development of the proposed phishing funnel model (PFM) artifact. PFM emphasizes the importance of the anti-phishing tool, phishing threat, and user-related factors in the decision-making process pertaining to four key funnel stages of the phishing attack: visit, browse, consider legitimate, and transaction. The model is estimated using a support vector ordinal regression with a custom kernel that parsimoniously captures users' funnel stage decisions across multiple phishing website encounters.

Design science research questions typically center on the efficacy of design elements within a proposed artifact (Abbasi et al. 2010) and how the artifact can "increase some measure of operational utility" (Gregor and Hevner 2013, p. 343). Accordingly, our research questions focus on predictive power and the downstream implications of better prediction.

RQ1. How effectively can PFM predict users' phishing susceptibility over time and in organizational settings?

RQ2. How effectively can interventions driven by susceptibility predictions improve avoidance outcomes in organizational settings?

To answer these questions, we evaluated PFM in two longitudinal field experiments. The first spanned a 12-month period within two organizations and involved 1,278 employees and 49,373 phishing interactions, highlighting PFM's ability to outperform competing models in predicting employees' susceptibility in real-world settings. The second was a follow-up three-month field study at the same two organizations examining the efficacy of interventions guided by susceptibility prediction; this follow-up experiment demonstrated the downstream value proposition of accurately predicting susceptibility.

From a design science perspective, PFM represents a novel solution (Gregor and Hevner 2013, Goes 2014). Although phishing is a known problem, *predicting user susceptibility* to phishing attacks is a new challenge that falls under the umbrella of proactive *security analytics*, which has been recently emphasized by various academics and practitioners (Chen et al. 2012, Musthaler 2013, Taylor 2014). Accordingly, the knowledge contributions of our work can be considered an *improvement*, based on recent design science guidelines (Gregor and Hevner 2013, Goes 2014). The proposed artifact and findings have implications for: (1) IT security managers tasked with real-time enterprise endpoint security and related organizational security policies and procedures and (2) Internet users in general.

This study addresses three important research gaps. First, prior work has not attempted to *predict* user susceptibility to phishing websites and has instead focused on developing or testing descriptive behavior models (Bravo-Lillo et al. 2011, Wang et al. 2012). The lack of predictive IT artifacts is a gap also noted by prior IS studies (Shmueli and Koppius 2011). We address this gap by not only demonstrating the feasibility of susceptibility prediction but also its efficacy as a potential component of real-time protection strategies. Second, prior phishing studies and user susceptibility models have typically focused on a single decision or action, such as considering a phishing website legitimate or being willing to transact with a phishing website (Grazioli and Jarvenpaa 2000, Dhamija et al. 2006, Sheng et al. 2010). However, falling prey to phishing website-based attacks entails a sequence of interrelated decisions and actions; modeling these sequences as a gestalt would thus provide deeper insight. Third, prior susceptibility models have

placed limited emphasis on anti-phishing tool and phishing threat-related factors despite their considerable impact on susceptibility to phishing attacks (Dhamija et al. 2006, Wu et al. 2006, Akhawe and Felt 2013).

2. Related Work

Traditionally, most of the research on anti-phishing has focused on benchmarking existing anti-phishing tools (Zhang et al. 2007, Abbasi et al. 2010) and developing better detection capabilities (Li and Schmitz 2009, Abbasi et al. 2010). Despite this research, phishing attacks have remained successful; thus, researchers and practitioners have increasingly turned their attention to user susceptibility. We define phishing susceptibility as the extent to which a user interacts with a given phishing attack. In recent years, several phishing susceptibility models have been proposed in an effort to *describe or explain* the salient factors attributable to users' susceptibility to phishing attacks (Downs et al. 2006, Bravo-Lillo et al. 2011).

The human-in-the-loop security framework (HITLSF) considers tool and user-related factors (Cranor 2008, Bravo-Lillo et al. 2011). Tool-related factors include whether the detection tool displays a warning, the user's level of trust in the tool, and the perceived usefulness of the tool's recommendations. User-related factors include demographics (e.g., age, gender, and education), knowledge (i.e., phishing awareness), prior experiences (e.g., past encounters/losses), and self-efficacy (i.e., ability to complete recommended actions). These factors impact the user's likelihood of visiting, browsing, and transacting with phishing websites (Bravo-Lillo et al. 2011).

Alnajim and Munro (2009) posited user-related technical abilities and phishing awareness as the two critical factors impacting users' decisions regarding the legitimacy of a particular website. When testing their model (which we refer to as AAM), they found that only awareness significantly impacted users' effectiveness in differentiating legitimate websites from phishing ones. Parrish et al. (2009) proposed a phishing susceptibility framework (PSF), which incorporates demographic factors (e.g., age and gender), experiential factors, big-five personality profile, and type of threat (e.g., the lure and hook in phishing emails). Sheng et al. (2010) investigated the impact of demographics, risk propensity, and knowledge of phishing on Internet users' ability to differentiate legitimate and phishing websites/emails (we refer to their model as DRKM). The demographic variables they used were age, gender, and education. Risk propensity implies a measure of willingness to engage in risky behavior. Knowledge and experience include phishing awareness, reliance on the web, and technical ability. Their analysis found that gender, age, and risk propensity significantly predicted users' ability to identify phishing threats.

Wang et al. (2012) developed a phishing susceptibility model (PSM) to explore threat and user-related factors in the context of phishing emails. Using a survey, they that found phishing knowledge, visceral cues, and deception indicators are the key drivers of participants' likelihood of responding to phishing emails. The PFM incorporates elements from each of these existing models while also introducing novelty in terms of independent variables incorporated, inclusion of multiple decision stages, and a parsimonious model estimation that considers user heterogeneity for predicting susceptibility.

3. The Phishing Funnel Model

Funnels have long been used to represent a series of interrelated decisions needed to accomplish a particular objective. In marketing, the awareness-interest-desire-action funnel for advertising dates back to the late 19th century (Jobber and Ellis-Chadwick 1995). The funnel shape represents attrition across stages: only a subset of decision makers at one stage of the funnel will continue on to the next. For instance, a particular advertisement will reach a subset of the target audience, a subset of those that view the advertisement will become interested, and an even smaller subset will actually make a purchase. In web analytics, conversion funnels are used to represent a website visitor's decision stages in e-commerce settings (Kaushik 2011). For example, a web conversion funnel for an e-tailer might entail the following stages: (1) visit the home page, (2) visit product pages, (3) add items to the shopping cart, (4) log in to the account, (5) proceed through checkout, and (6) receive an order confirmation.

The funnel concept is also highly relevant for modeling phishing. Users typically encounter a phishing attack in one of the following ways: (1) through a phishing email containing a uniform resource locator (URL) to a website (Wright and Marett 2010; Hong 2012; Wang et al. 2012, 2016); (2) through search engine results, where fraudulent websites often rank highly using black-hat search engine optimization (Gyongyi and Garcia-Molina 2005); or (3) through social media, including blogs, forum postings, comments, tweets, and so on (Kolari et al. 2006). Regardless of how phishing sites are initially encountered, users are faced with four progressively dangerous decisions that determine their susceptibility. First, users must decide whether to click on the link to visit the website (Jagatic et al. 2007). Second, those that visit must decide whether to browse the website, where browsing is typically defined in terms of engagement with the site, such as the amount of time spent viewing a page or the quantity of pages viewed (Bravo-Lillo et al. 2011,

Kaushik 2011). Third, users that browse must deem the site legitimate before considering engaging in transactions (Alnajim and Munro 2009). Fourth, users must decide whether to transact with the website, which can result in identity theft and monetary losses (Grazioli and Jarvenpaa 2000, Abbasi et al. 2010). Users do not need to reach the final stage to be exposed to fraud and security risks; for example, simply visiting or browsing can expose users to malware (Bravo-Lillo et al. 2011, Verizon 2016). Scammers hope to entice as many unsuspecting users as far down the funnel as possible, thereby giving the funnel a wide cylindrical shape; by contrast, the ideal scenario from a user's perspective is to avoid the funnel entirely.

Figure 1 shows the PFM, a design artifact for predicting user susceptibility to phishing websites. PFM encompasses six categories of factors that impact decision-making related to phishing susceptibility (top left of the figure). The tool, threat, and user susceptibility factors are used as independent variables to predict user susceptibility (where the funnel stages on the top right signify the dependent variable). Susceptibility is predicted as an ordinal response indicating the final funnel stage for a given user-phish encounter. The predictive model is operationalized via a support vector ordinal regression (SVOR) method that incorporates a custom kernel function that uses a cumulative link mixed model (CLMM). Having already described the funnel concept, in the remainder of the section, we elaborate on the susceptibility factors and support vector ordinal regression method.

3.1. Susceptibility Factors Incorporated in PFM

PFM encompasses six categories of factors that impact decision-making related to phishing susceptibility. These factors pertain to: (1) the tool, (2) the threat, and (3) characteristics of the user. Because no single theoretical framework incorporates all three of these factors, we draw from three primary theories: (1) the technology acceptance model (TAM; Davis 1989), protection motivation theory (PMT; Rogers and Prentice-Dunn 1997), and the human-in-the-loop literature (Cranor 2008, Kumaraguru et al. 2010). We describe how each of these theories/bodies of knowledge (summarized in Table 1) informs our selection of variables.

3.1.1. Tool Factors and the TAM. As explained by TAM, the adoption of and reliance on an anti0-phishing tool depend on perceptions of both its usefulness and its ease of use. These two factors have significantly predicted adoption in a wide variety of applications and contexts (Benbasat and Barki 2007), including anti-phishing tools (Herath et al. 2014) and security tools generally (Kumar et al. 2008). Accordingly, in addition to collecting objective measures of performance of the antiphishing tools (i.e., tool

Figure 1. Phishing Funnel Model

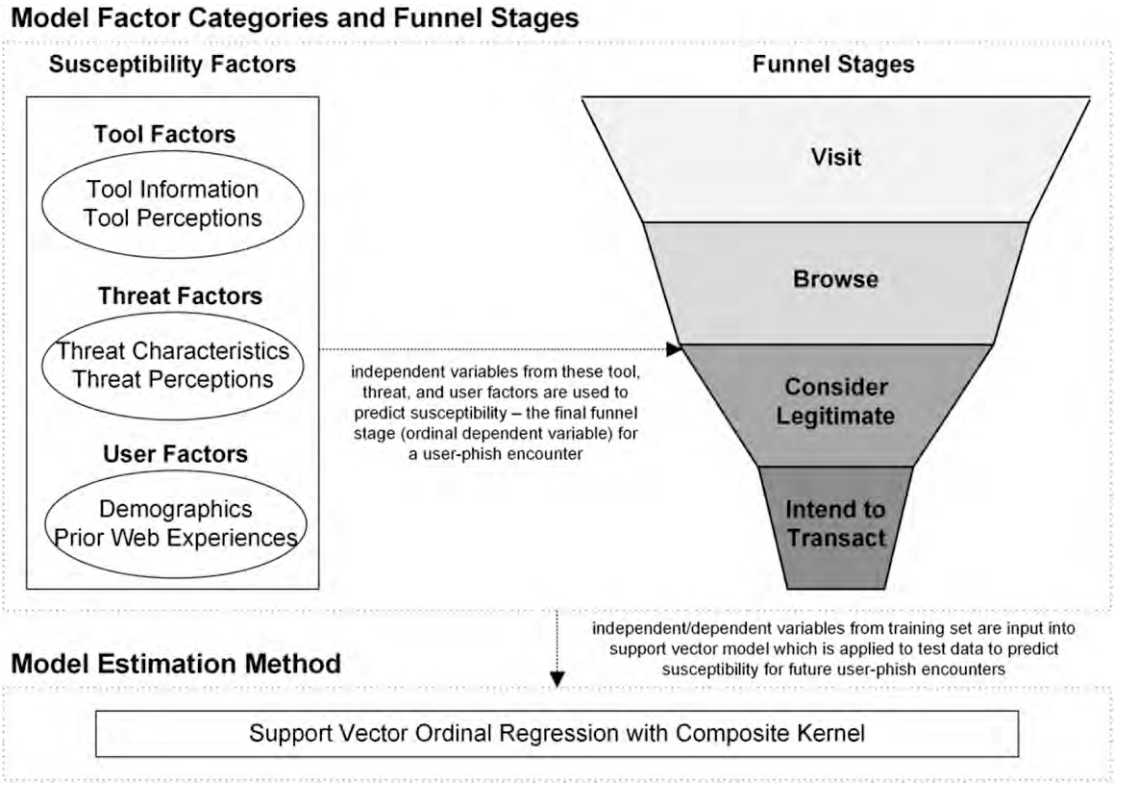


Table 1. Variables Related to Categories of Susceptibility Factors in PFM and Their Mapping to Theoretical Constructs

PFM factors and subcategories		Theory and constructs		PFM variables	References	Application of theory to PFM
Tool factors	Tool information	Technology acceptance model	Perceived usefulness	Tool warning	Wu et al. 2006; Cranor 2008; Bravo-Lillo et al. 2011	The adoption of and reliance on an antiphishing tool depend on perceptions of its usefulness and ease of use.
				Tool detection rate	Abbasi et al. 2010; Hong 2012	
				Processing time	Dhamija et al. 2006	
	Tool perceptions			Tool usefulness	Venkatesh et al. 2003; Cranor 2008; Egelman et al. 2008	
			Perceived ease of use	Tool effort required	Davis 1989; Venkatesh et al. 2003; Keith et al. 2009	
				Cost of tool error	Cavusoglu et al. 2005; Liang and Xue 2009	
Threat factors	Threat characteristics	Protection motivation theory	Prior threat experiences	Threat domain	Grazioli and Jarvenpaa 2003; Bansal et al. 2010; Angst and Agarwal 2009	Responses to threats depend on perceptions of threat severity and susceptibility, informed by prior experience.
				Threat type	Dhamija et al. 2006; Parrish et al. 2009	
				Threat context	Lennon 2011; McAfee 2013	
			Threat severity	Threat severity	Kaushik 2011; Ma et al. 2012; Vishwanath et al. 2011; Bar-Ilan et al. 2009; Wang et al. 2011; Agarwal et al. 2011	
	Threat perceptions			Phishing awareness	Downs et al. 2006; Alnajim and Munro 2009; Bravo-Lillo et al. 2011; Wang et al. 2012; Wang et al. 2016	
			Threat susceptibility	Perceived severity	Downs et al. 2007; Camp 2009; Liang and Xue 2009; Zahedi et al. 2015; Wang et al. 2017	
User factors	Demographics	Human in the loop	Demographics	Gender	Venkatesh et al. 2003; Morris et al. 2005; Jagatic et al. 2007; Sheng et al. 2010	Demographics, personal characteristics, and knowledge and experience, influence warning effectiveness.
				Age	Venkatesh et al. 2003; Cranor 2008; Parrish et al. 2009; Sheng et al. 2010	
			Personal characteristics	Education	Porter and Donthu 2006; Sheng et al. 2010	
	Prior web experiences		Knowledge and experience	Trust in institution	Pavlou and Gefen 2004	
				Familiarity with domain	Kumaraguru et al. 2010	
				Familiarity with site	Dhamija et al. 2006; Wu et al. 2006; Kumaraguru et al. 2010	
				Past losses	Downs et al. 2006	

warning, detection rate, and processing time), we also capture users' perceptions of the tool's usefulness and effort required to use (i.e., ease of use). Additionally, we captured the cost of tool error, a variable that adversely affects ease of use (Cavusoglu et al. 2005, Liang and Xue 2009). Consistent with TAM, users' reliance on the anti-phishing tool should depend on perceptions of usefulness, the effort required, and the cost of tool error.

3.1.2. Tool Factors—Tool Information. Tool information variables include *tool warnings*, *detection rates*, and *processing times*. Once a user enters a URL or clicks on a link, the anti-phishing tool determines whether the website associated with the URL poses a threat (Zhang et al. 2007, Hong 2012). For URLs deemed to be potential phishing sites, users encounter a warning page designed to dissuade them from proceeding to the initial visit phase of the phishing funnel; alternatively, for websites deemed legitimate, no warning is presented. The presence or absence of this warning can significantly impact users' decisions and actions regarding various funnel stages. For example, the presence of a warning may reduce the likelihood of visiting a website or of browsing a website that has already been visited (Bravo-Lillo et al. 2011). Warnings may also affect perceptions regarding the legitimacy of a website (Wu et al. 2006, Cranor 2008).

For tools to display a meaningful warning, they must be capable of accurate detection of potential phishing sites; benchmarking studies have shown that typical detection rates are between 60% and 90% (Zhang et al. 2007, Abbasi et al. 2010, Hong 2012). Lack of adequate detection rates can cause users to disregard tool recommendations (Sunshine et al. 2009). Moreover, benchmarking studies have also found that tool processing times typically range from one to four seconds (Abbasi et al. 2010). Because users consider security warnings a secondary task that distracts from their primary objective (Dhamija et al. 2006, Jenkins et al. 2016), processing times may impact how users react to tool recommendations.

3.1.3. Tool Factors—Tool Perceptions. The IS literature examining users' perceptions of various technology tools has identified a core set of constructs that predict individual use of technologies (Venkatesh et al. 2003). Within that set, perceptions of a given technology's *usefulness* are often the strongest predictor of system use in most settings (Venkatesh et al. 2003). Perceived usefulness has also been theorized as a predictor of anti-phishing tool use (Cranor 2008). Users with low perceived usefulness of anti-phishing tools may ignore tool warnings, thereby increasing susceptibility (Egelman et al. 2008).

In addition to tool usefulness, user perception of effort has been a strong predictor of system use (Davis 1989, Venkatesh et al. 2003). User tasks associated with anti-phishing tools include waiting for the tool to evaluate a clicked/typed URL, reading tool warnings, and deciding whether to adhere to tool recommendations. Although *tool effort required* has not been included in existing phishing susceptibility models, it has been incorporated in studies on other security problems (Keith et al. 2009).

Finally, the perceived *cost of a tool error*, defined as the perceived cost of following an incorrect recommendation, is a key determinant of tool use. The most common and severe form of classification error for anti-phishing tools is a false negative or classifying a phishing website as legitimate (Zhang et al. 2007, Akhawe and Felt 2013). False negatives prevent proper security warnings and thereby increase susceptibility to phishing attacks, resulting in monetary consequences (Cavusoglu et al. 2005). Such failures impact users' cost-benefit evaluation regarding threat countermeasures (e.g., detection tools; Liang and Xue 2009), which could hinder tool use. However, perceptions of false positives can also lead to the *cry wolf* effect, causing users to discount future tool warnings (Sunshine et al. 2009). Furthermore, perceived costs of tool error may not be entirely correlated with actual tool errors and costs, with some users perceiving such costs to be much higher than others (Zahedi et al. 2015).

3.1.4. Threat Factors and PMT. PMT is widely used in IS to explain security-related behaviors (Liang and Xue 2009, Boss et al. 2015, Cram et al. 2019). At the core of PMT are two cognitive mediating processes that occur when a person encounters a threat: threat appraisal and coping appraisal (Floyd et al. 2000). Threat appraisal involves assessing both the severity of a threat and one's vulnerability to it. At the same time, the coping appraisal process evaluates the effectiveness of possible responses and one's own ability to enact those responses. Importantly, both these processes are influenced by information about the environment and one's prior experience (Rogers and Prentice-Dunn 1997). Accordingly, we capture variables relating to the threat severity of phishing and users' perceptions of these threats. Additionally, following PMT, we also include variables relating to the domain, context, and users' awareness of phishing threats informed by their own threat experiences. In line with PMT, users' susceptibility to traversing the phishing funnel stages will be predicted by these threat factors.

3.1.5. Threat Factors: Threat Characteristics. *Threat domains* include e-commerce platforms such as

business-to-customer and business-to-business platforms (Grazioli and Jarvenpaa 2003) and industry sectors such as financial, health, retail, and so on (Abbasi et al. 2010). Threat domains can impact users' intentions to disclose personal information (Bansal et al. 2010), thereby influencing susceptibility to phishing attacks. In highly sensitive domains such as finance and health, users may be more risk averse (Angst and Agarwal 2009).

The phishing *threat type* a user is exposed to can impact the likelihood of susceptibility (Parrish et al. 2009, Wright et al. 2014). Dhamija et al. (2006) found that certain threat types had success rates that were orders of magnitude higher than other attacks. Two common types of phishing threats are *concocted* and *spoof* websites. Concocted websites seek to appear as unique, legitimate commercial entities in order to engage in failure-to-ship fraud (i.e., accepting payment without providing the agreed upon goods/services) and often rely on social engineering-based attacks to reach their target audience (Abbasi et al. 2010). For instance, fraudulent eBay sellers may gain buyers' trust by going through a seller-controlled concocted online escrow website (Chua and Wareham 2004, Abbasi et al. 2010). Conversely, spoof websites engage in identity theft by mimicking legitimate websites to target users familiar with the legitimate website and brand (Dhamija et al. 2006, Dinev 2006, Liu et al. 2006).

Threat severity must also be considered, given that users tend to be more risk averse when stakes are higher (Kahneman and Tversky 1979, Zahedi et al. 2015). Prior work has found that the median losses attributable to phishing range from approximately $300 for those suffering only direct monetary losses to $3,000 for victims of identity theft, with the latter amount including remediation and reputation costs (Lennon 2011, McAfee 2013). Threats that are more severe in terms of potential losses are likely to garner more conservative user behavior with respect to funnel-related decisions (Zahedi et al. 2015).

Threat context factors can also impact users' perceptions, decisions, and actions in online settings. For instance, a user's email load can impact his or her response rate to phishing email-based attacks (Vishwanath et al. 2011). For search engines, click-through rates and user trust are higher for web pages that are ranked higher in search results (Bar-Ilan et al. 2009, Kaushik 2011, Ma et al. 2012), which in turn leads to online scammers expending effort to influence search result placement (Wang et al. 2011).

3.1.6. Threat Factors: Threat Perceptions. When encountering a potential phishing attack, users' perceptions of the threat and their resulting judgments are key prerequisite considerations for any decisions and actions (Bravo-Lillo et al. 2011). Greater *perceived phishing severity* is likely to result in greater protective behavior (Camp 2009, Zahedi et al. 2015). For example, Downs et al. (2007) observed that users who indicated a higher perceived threat severity for having their information stolen were less likely to transact with potential phishing websites.

Awareness of phishing attacks is another critical factor impacting users' decisions and actions in various phishing funnel stages. People with greater *phishing awareness* are likely to be more knowledgeable about the threat and hence capable of making better decisions (Bravo-Lillo et al. 2011, Wang et al. 2012). For instance, Downs et al. (2006) found that users with greater self-reported phishing awareness viewed the consequences of phishing attacks differently than those with less awareness, and Alnajim and Munro (2009) showed that users with greater phishing awareness were less likely to consider a phishing website legitimate.

3.1.7. User Factors and the Human-in-the-Loop Literature. In addition to tool and threat factors, the characteristics of users themselves are also theorized as substantially influencing decisions to heed security warnings (Anderson et al. 2016a, b). An inclusive theoretical framework describing this process from the human-computer interaction literature is the human-in-the-loop security framework (HITLSF). The HITLSF and DRKM models adopted as benchmarks in our study (Cranor 2008, Sheng et al. 2010, Bravo-Lillo et al. 2011) belong to this body of literature. HITLSF explains that demographics such as *age*, *gender*, and *education* can substantially mediate the effectiveness of warnings on security behavior. We therefore capture these variables in PFM. Similarly, related studies that have espoused the HITLSF perspective hold that knowledge and experience also mediate the effectiveness of warnings (Dhamija et al. 2006, Downs et al. 2006, Kumaraguru et al. 2010, Sheng et al. 2010). We likewise include in PFM the variables of *familiarity of domain*, *familiarity with site*, and *past losses*, the latter of which has been shown to be especially important to users' decisions to heed security warnings (Vance et al. 2014).

Finally, a key factor derived from past experience is trust in an institution (McKnight et al. 1998, Pavlou and Gefen 2004). Trust, by definition, is a willingness to become vulnerable to someone or something (Mayer et al. 1995) and is foundational to a range of online behaviors (McKnight et al. 2002). Phishing effectively exploits users' trust in familiar institutions with which they are accustomed to interacting

(Oliveira et al. 2017). Therefore, consistent with HITLSF, we capture *trust in institutions* as an important aspect of past experience.

3.1.8. User Factors: Demographics. Among an almost limitless range of demographic variables that could potentially influence technology use, only a relative few have consistently proven to significantly influence if, when, or how technologies are used and decisions are made. Foremost among these is perhaps *gender* (Gefen and Straub 1997). Research has shown that men tend to focus on instrumental outcomes, whereas women use a more balanced or holistic set of criteria in evaluating potential use (Morris et al. 2005). In prior phishing susceptibility studies, gender has been found to be a significant factor (Parrish et al. 2009, Sheng et al. 2010).

Age has also been shown to exert an important influence on technology adoption and use (Morris et al. 2005), and prior phishing susceptibility studies have identified age as an important factor (Cranor 2008, Parrish et al. 2009). For instance, Sheng et al. (2010) found age to be significant, with younger adults exhibiting greater susceptibility. Similarly, prior studies have demonstrated that *education* has a differential effect on adoption and use (Porter and Donthu 2006). In the phishing context, education may be correlated with technical training and knowledge, which can impact phishing susceptibility (Sheng et al. 2010).

3.1.9. User Factors: Prior Web Experiences. Experience-related variables can have profound and complex effects on users' decisions and actions. *Trust in institution* has been shown to be an important factor impacting users' online decisions (Pavlou and Gefen 2004). Users that are more trusting of banking websites in general are far more likely to use their bank's online services (Freed 2011). Similarly, users that are more trusting of health infomediaries are more likely to use services offered by specific online health resources (Zahedi and Song 2008).

Familiarity with websites may have different effects on user susceptibility to phishing attacks (Kumaraguru et al. 2010). Although website familiarity may help detect phishing in some situations, it can also be exploited by certain types of phishing attacks (Dinev 2006); for example, a user familiar with a particular website may be fooled by visual deception attacks (Dhamija et al. 2006). In addition, Wu et al. (2006, p. 606) found that many users incorrectly considered phishing websites legitimate because the web content looked "similar to what they had seen before." *Familiarity with a domain* such as online banks or online pharmacies might similarly affect users' perceptions (Kumaraguru et al. 2010).

Past losses resulting from exposure to phishing websites can influence users' decisions and actions pertaining to current/future phishing funnel stages. One would assume that the *fool me twice, shame on me* logic applies. However, Downs et al. (2006) found that users who had experienced prior losses were more than 50% more likely to fall prey to a phishing attack and they attributed this finding to a possible inherent *gullibility* to phishing attacks among users.

3.2. Prediction Using Support Vector Ordinal Regression with CLMM

The phishing funnel involves four binary decision stages, each of which could be treated as a separate binary classification problem. However, such an approach would present challenges emerging from cross-stage interdependencies. Because of theoretical and statistical considerations guided by model parsimony, we treat the funnel as a single ordinal response variable with five possible end outcomes: *no visit, visit, browse, consider legitimate,* and *intend to transact*, which we model as an ordinal regression. The five possible phishing funnel end points could be modeled using equidistant threshold values, thereby simplifying the ordinal models (Shashua and Levin 2003, Christensen 2015). However, progression through funnel stages does not necessarily occur in equally sized steps. For example, it is highly plausible that the choice to stop at *browse* rather than at *visit* is more commonplace than proceeding past *browse* to *consider legitimate*. Even in marketing conversion funnels, abandonment rates have been shown to be higher at select stages because of users' perceptions that these stages entail *bigger decisions* (Kaushik 2011). Hence, we use ordinal regression models with flexible, nonequidistant thresholds.

Kernel-based machine learning methods have been used by IS researchers in recent years based on their ability to derive patterns from noisy data and incorporate theory-driven design (Abbasi et al. 2010). By using the *kernel trick*—representing all N instances in the training data as a positive semidefinite, symmetric $N \times N$ matrix—such methods are able to incorporate nonlinear domain-specific functions into a linear learning environment (Burges 1998). In our context, they afford opportunities to incorporate custom kernel functions that capture key elements of PFM, such as user, tool, and threat-related susceptibility predictors, interrelated funnel stages, and flexible cross-stage thresholds. Accordingly, we propose a support vector ordinal regression (Chu and Keerthi 2007) with a composite kernel (SVORCK). Our composite kernel function, K_{PFM} is

$$K_{PFM} = K_{UTT} + K_{Funnel}, \quad (1)$$

where K_{UTT} is a linear kernel that takes the user, tool, and threat variables as input for any two user-phish encounters g and h, and applies a dot-product transformation between their respective feature vectors $\mathbf{a}_g$ and $\mathbf{a}_h$:

$$K_{UTT}(g,h) = \frac{\langle \mathrm{a}_g, \mathrm{a}_h \rangle}{\sqrt{\langle \mathrm{a}_g, \mathrm{a}_g \rangle \langle \mathrm{a}_h, \mathrm{a}_h \rangle}}. \tag{2}$$

Whereas K_{UTT} addresses user, tool, and threat considerations associated with the observe and orient stages in PFM, the funnel kernel K_{Funnel} takes into account funnel stage traversal information associated with the decide and act stages of PFM while also considering user effects. For a given user i, let $j = 1, \ldots, n_i$ denote the set of user-phish encounters associated with that i (i.e., repeated measures). Let $c = 1, 2, \ldots, C$ represent the response categories, which in this case represent final funnel stage categories such as no-visit, visit, browse, consider legit, and intend to transact. Then, Y_{ij} is the ordinal response associated with user i and user-phish encounter j. The funnel kernel, K_{Funnel}, runs a cumulative-link mixed model over the user, tool, and threat variables to produce a vector of funnel stage probabilities for each user-phish encounter, $\mathbf{d}_{\mathrm{ij}}$. A key benefit of the inclusion of the CLMM in our SVORCK is its ability to measure funnel stage traversal in a manner that accounts for user effects via the mixed model. We define the cumulative probabilities for the C categories of our ordinal funnel outcome Y as

$$P_{ijc} = \Pr(Y_{ij} \le c) = \sum_{k=1}^{c} p_{ijk}, \tag{3}$$

where p_{ijk} represents the individual category probabilities. The CLMM is represented as

$$\lambda_{ijc} = \log\left[\frac{p_{ijc}}{1 - p_{ijc}}\right] = \gamma_c - [x'_{ij}\beta + z'_{ij}T\theta_i] \tag{4}$$

for $c = 1, \ldots, C-1$, where $\boldsymbol{x_{ij}}$ is the covariate vector, $\boldsymbol{\beta}$ is the regression parameter vector, and $\boldsymbol{z_{ij}}$ is the vector of random-effect variables. The random effects follow a multivariate Gaussian distribution with variance-covariance matrix Σ_v and mean vector 0—we standardize these to $T\theta_i$, where $TT' = \Sigma_v$ is the Cholesky decomposition, and θ_i follows a standard multivariate normal distribution. γ_c is one of the $C - 1$ thresholds such that $\gamma_1 < \gamma_2 \ldots < \gamma_{C\text{-}1}$. Because of the proportional odds assumption (McCullagh 1980), the regression coefficients β do not include the c subscript. Using the CLMM output, each user-phish encounter can be represented as a vector of funnel traversal probabilities: $\mathbf{d}_{\mathrm{ij}} = (\lambda_{ij1}, \lambda_{ij2,\ldots,}\lambda_{ijC})$.

The funnel kernel, K_{Funnel}, can compare funnel traversal probabilities between any two user-phish instances g and h, once again using a dot-product transformation between their respective CLMM-based funnel probability vectors $\mathbf{b}_g$ and $\mathbf{b}_h$:

$$K_{Funnel}(g,h) = \frac{\langle \mathrm{b}_g, \mathrm{b}_h \rangle}{\sqrt{\langle \mathrm{b}_g, \mathrm{b}_g \rangle \langle \mathrm{b}_h, \mathrm{b}_h \rangle}}, \tag{5}$$

where each g and h maps to a specific ij, and consequently each $\mathbf{b}_g$ and $\mathbf{b}_h$ equals some $\mathbf{d}_{\mathrm{ij}}$. Finally, our composite kernel K_{PFM}, which combines K_{UTT} and K_{Funnel}, can be computed as follows:

$$K_{PFM}(\mathrm{a}_g + \mathrm{b}_g, \mathrm{a}_h + \mathrm{b}_h) = \frac{\langle \mathrm{a}_g, \mathrm{a}_h \rangle}{\sqrt{\langle \mathrm{a}_g, \mathrm{a}_g \rangle \langle \mathrm{a}_h, \mathrm{a}_h \rangle}} + \frac{\langle \mathrm{b}_g, \mathrm{b}_h \rangle}{\sqrt{\langle \mathrm{b}_g, \mathrm{b}_g \rangle \langle \mathrm{b}_h, \mathrm{b}_h \rangle}}. \tag{6}$$

In the ensuing experiments, we report the results for PFM using both the SVORCK and CLMM. We show that PFM-CLMM outperforms comparison methods, while SVORCK offers further significantly enhanced predictive power relative to CLMM.

4. Evaluation

To address our research questions, we conducted two longitudinal field experiments, summarized in Table 2. For RQ1, we conducted a longitudinal field experiment over the course of 12 months to test the ability of PFM to predict the phishing susceptibility of employees at two organizations. For RQ2, we followed up our prediction field experiment with a three-month field study to test the value of interventions guided by susceptibility prediction.

5. Experiment 1: Prediction—Field Testing PFM Longitudinally in Two Organizations

To answer RQ1, we conducted a longitudinal field experiment that examined phishing susceptibility behavior and intentions. A longitudinal design was used to account for changes in participants' perceptions of new web experiences, encounters with threats, and interactions with antiphishing tools.

Experiment 1 was performed within two organizations: a large financial services company (FinOrg) and a midsized legal services firm (LegOrg). In each organization, employees with access to work-related computers were invited by high-level executives to participate in the experiment. Employees were not given details about the nature or purpose of the study—they were simply told that they would be asked to respond to quarterly surveys and periodically answer pop-up questions. In both companies,

Table 2. Summary of Experiments

Research question	Experiment type/ duration	Participants (employees at FinOrg and LegOrg)	Data points	Final dependent variables
RQ1. How effectively can PFM predict users' phishing susceptibility over time and in organizational settings?	Prediction: Longitudinal (12 months)	1,278	49,373	(1) Intention to transact with phishing website; (2) Observed transacting behavior
RQ2. How effectively can interventions driven by susceptibility predictions improve avoidance outcomes in organizational settings?	Intervention: Longitudinal (3 months)	1,218	13,824	

management incentivized employee participation by offering additional paid time off commensurate with participation duration. Table 3 provides an overview of the study participants; during the study's 12-month period, 50 participants (~4%) dropped out mostly because of normal turnover.

As a precursor to the field experiment, we conducted two preliminary, laboratory-based experiments to pretest the proposed PFM predictive model. These laboratory experiments were conducted in a university setting and then repeated with individual B2C customer of a major security software provider. The results were used to validate our choice of susceptibility predictors, survey items, and operationalizations for PFM and comparison methods. Online Appendix A lists the final PFM survey instrument for various tool, threat, and user construct variables incorporated into the model; moreover, we included appropriate items pertaining to PFM's competitor models as noted in Online Appendix C.

5.1. Experiment 1: Prediction—Design

During the field experiment, all of the work computers of FinOrg participants were equipped with an enterprise endpoint security solution capable of detecting email and web-based phishing threats using robust rule-based and machine learning-driven analysis of URLs and website content. This solution used client-side servers coupled with a third-party enterprise security provider's machine-learning servers. Similarly, for the duration of the field experiment, LegOrg participants' work computers were equipped with an endpoint protection solution designed for small- to medium-sized businesses. This offered a more nimble solution that did not require constant interaction with the third party provider's servers. The detection rates and processing times for the FinOrg and LegOrg anti-phishing tools are provided in Table 4. Both software packages displayed prominent warnings whenever a URL deemed to be a potential phish was clicked on.

It is worth noting that measuring threat characteristic variables in real-time field settings entails mechanisms for identifying threat domain, the potential type of threat, and potential severity of a threat. As noted in Table 4, we used algorithms capable of accurately inferring the domain and potential type of a website. Similarly, whether a given URL or web session exposes a user to malware is a well-studied problem (Rajab et al. 2011). However, the variable measurements are not perfect, as the threat domain, type, and severity classification methods do produce errors (albeit in a small proportion of cases). Because the field experiment occurred in real time as participants interacted with websites on their work computers, a mechanism was necessary to collect funnel stage variables from all *potential* phishing websites, irrespective of whether the website had been verified as phishing or not. A URL appearing on a user's screen as part of an email, search result, link in a web page or social media post, and so on, was operationalized as a potential phish if (1) the organizations' endpoint security tool considered it to be a phish, in which case a warning would appear or (2) the URL appeared in any of several reputable phishing website databases as either verified or pending based on a real-time check.

Funnel stages were also determined for each potential phishing URL. Visitation and browsing decisions were automatically recorded from clickstream logs.

Table 3. Overview of Field Study Participants in Experiment 1

Company	Industry	Company size	No. invited	No. participants	Opt-in rate	Average age (years)	Gender (female)	Bachelor degree
FinOrg	Financial	Large	1,151	796	69.2%	34.1	30.0%	90.1%
LegOrg	Legal	Midsized	655	482	73.6%	37.6	48.9%	86.5%
Total			1,806	1278	70.8%	35.4	37.2%	88.7%

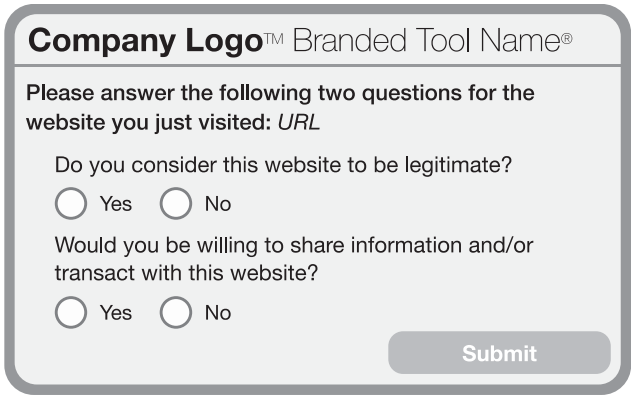
Table 4. Operationalization of Select Field Study Variables

Category	Variable	Description
Tool information	Tool detection rate	FinOrg's tool's rated detection rate was 98%, although FinOrg's IT security staff indicated an observed rate of 96% during an extended period prior to the field study. LegOrg tool's observed rate was 87% based on an analysis of historical system logs.
	Tool warning	Whether a warning was displayed for that given URL (1 = warning; 0 = no warning).
	Tool processing time	FinOrg's tool had a mean run time of 0.9 seconds; LegOrg's tool had a mean run time of 1.9 seconds.
Threat characteristics	Threat domain and threat type	Seven domains: financial services, retail, information, professional services, transportation, entertainment, and health. Two threat types: concocted and spoof. Threat domain and type were computed by comparing the similarity of each potential phishing site against a database of thousands of prior known phishing sites catalogued with their accompanying threat domain and type labels. Similarity assessment algorithms have been shown to accurately determine phishing site domain (e.g., finance, entertainment) and threat type (e.g., spoof or concocted; Liu et al. 2006, Qi and Davison 2009).
	Threat severity	Two settings: high and low. Websites with malware, as determined using FinOrg and LegOrg's enterprise web malware detection, were categorized as "high severity" since this posed additional threat atop the inherent identity theft risk.
	Threat context	Ranging from 1 to 10, where lower values indicate greater primacy. For URLs appearing in search engine results, order was the search result ranking. For URLs appearing in emails, order was an ascending percentile rank across all newly received emails. For instance, if the URL appeared as the 3rd of 5 new emails, the order would be 6 (i.e., 3/5 = 6/10). A similar ascending percentile rank conversion was used for URLs appearing in social media comments (e.g., Facebook).
Demographics	Age, gender, education	The age, gender, and education level of each employee (provided by the organizations). Education levels ranged from high school graduate to doctoral degree.
Prior web experiences	Trust in institution and familiarity with domain	Using North American Industry Classification System (NAICS) guidelines, participants rated their familiarity and trust with various website domains including financial services, retail, information, professional services, transportation, entertainment, and health.
	Familiarity with site	Participants rated their familiarity with 200 websites commonly targeted by phishing attacks compiled from (1) various databases such as PhishTank and the Anti-Phishing Working Group and (2) drawn from an analysis of URLs in the two organizations' Internet usage logs.

A *visit* was recorded when the user explicitly clicked on the URL and arrived on the phishing site's landing page. When presented with a tool warning, this involved circumventing the warning by clicking the option to continue to the site. Following the web analytics literature (Kaushik 2011), a *browse* was recorded when a user either clicked on a link while on the site or spent at least 30 seconds on the landing page (as the active browser window). Once participants concluded sessions with a potential phishing site, a pop-up form asked if they *considered the site legitimate* and/or *intended to transact* with the site. Figure 2 shows an illustration of the pop-up form. Although these questions were asked for all potential user-phish encounters, they contributed to determining the final funnel stage only for sessions in which the user actually visited and browsed the site. *Observed transactions* were also recorded.

For the purposes of prediction, the field experiment used a windowed approach as shown at the bottom of Figure 3: for example, within the first window, months 1–3 were used for training and months 4–6 were used for testing; in the following window, months 4–6 were used for training, whereas months 7–9 were used for testing, and so on. Before each window (e.g., before the start of months 1–3), surveys were used to gather participants' tool perception, threat perception, user experiences, and demographic information for PFM, as well as the items necessary for HITLSF, DRKM, and AAM. The timing of these longitudinal surveys is indicated at the top of Figure 3. Additional details regarding the operationalization of the PFM non–survey-based variables and the familiarity survey items appear in Table 4. As noted, survey-based item details can be found in Online Appendix A. To ensure survey

Figure 2. Illustration of the Pop-Up Form



Note. This form was displayed to participants at the end of each session with a potential phishing site.

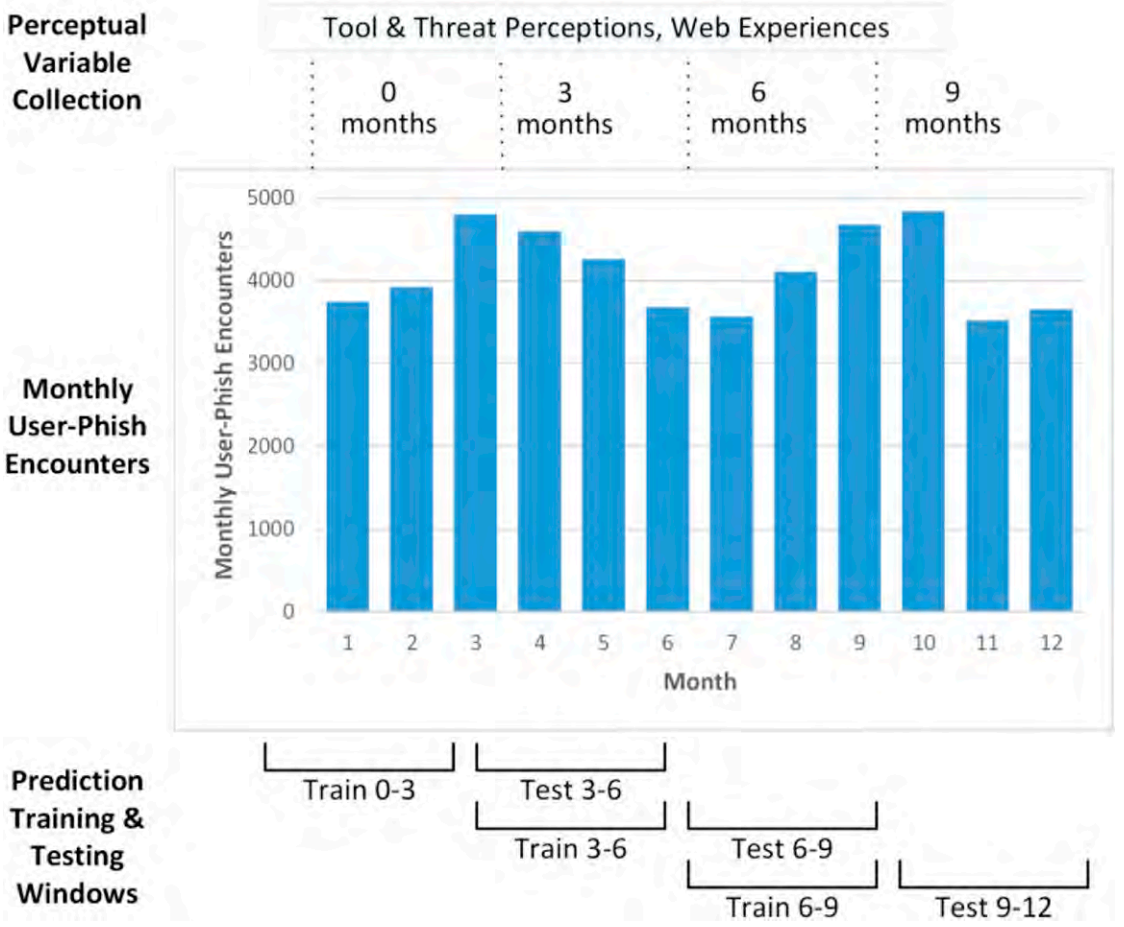
construct reliability and convergent and discriminant validity for the survey items incorporated in PFM, we performed a series of analyses on the first (i.e., month 0) survey data collection (see Online Appendix B). Exploratory factor analysis showed that for a given construct, all associated survey items loaded on the same factor. Additionally, Cronbach's alpha values were computed to ensure construct reliability. Consistent with prior work, we ultimately averaged survey items to arrive at a single value per construct. None of the constructs were highly correlated.

All potential phishing URLs encountered by the 1,278 participants during the entire 12-month period were eventually verified against online databases, resulting in a test bed of 49,373 verified participant-phish encounters. As depicted using the bar chart in Figure 3, this averaged out to 4,100 mean monthly participant-phish encounter instances (~3.25 URLs per participant per month). Summary statistics for all PFM susceptibility independent variables, across the 12-month period, appear in Online Appendix B.

5.2. Experiment 1: Prediction—Results

Two analyses were conducted. In the first, we evaluated the predictive power of PFM relative to the competing DRKM, AAM, and HITLSF. Each of the three competing models were trained using CLMM

Figure 3. (Color online) Illustration of 12-Month Field Experiment Design



Note. Top shows quarterly survey timing for perceptual variable collection; middle shows monthly user-phish encounters across the two organizations; bottom depicts the training/testing windows for all models.

with flexible thresholds, allowing for apples-to-apples comparison of the different combinations of independent variables across these models. Moreover, in addition to the PFM model using our proposed SVORCK method, we evaluated an additional CLMM model trained without the composite kernel to assess the additive value of the composite kernel.

In the second analysis, we compared PFM with existing benchmark methods for behavior prediction using the same set of PFM variables: these methods included Bayesian network (BayesNet) and support vector machines (SVMs)—which have been previously used for behavior prediction—and basic SVOR, a CLMM variant with equidistant thresholds, and a linear mixed model (LMM) baseline.

Given that predicting users' end funnel stages is an imbalanced multiclass classification problem, we used multiclass receiver operating characteristic (ROC) curves and area-under-the-curve values (AUC) to assess predictive model tradeoffs between true/false positives (Fawcett 2006, Bardhan et al. 2015). The use of these measures is consistent with prior design science studies pertaining to predictive artifacts (Prat et al. 2015). All models and methods were evaluated on the 36,909 test instances that transpired over the last nine months (i.e., months 4–12).

As shown in Table 5, PFM—using SVORCK or CLMM—significantly outperformed the three comparison models with AUC values that were 22%–35% higher, and PFM's AUC was also between 8% and 25% higher than the competing susceptibility prediction methods (all $p < 0.001$). Figure 4 shows the accompanying ROC curves depicting model tradeoffs between true (y axis) and false (x axis) positive rates. As illustrated, both PFMs' ROC curves outperformed their peers with markedly higher true-positive rates for most levels of false positives. Within PFM, SVORCK once again yielded a four-percentage-point lift over CLMM ($p < 0.001$). When garnering 90% true positives, PFM-SVORCK had a false-positive rate of about 33%, whereas PFM-CLMM had a 40% rate, and the best comparison models and methods attained false-positive rates of around 70%. Collectively, these results show that both the choice of dependent variables and the methods used have a substantial impact on predicting phishing susceptibility, with the former having slightly more impact, as observed by differences in AUC.

To illustrate the utility and practical significance of PFM's predictive performance lift for FinOrg and LegOrg, we examined the phishing funnel across the 12-month field experiment. The observed funnel stage traversal frequencies (left chart) and percentages (right funnel) are depicted in Figure 5. We found that 3.8% of employees' participant-phish encounters resulted in an intention to transact, equating to 1,896 total instances across the two organizations over the entire 12-month period, and found that employees visited over 50% of the phishing websites encountered, including 3,216 URLs deemed to be high severity (i.e., containing potential malware).

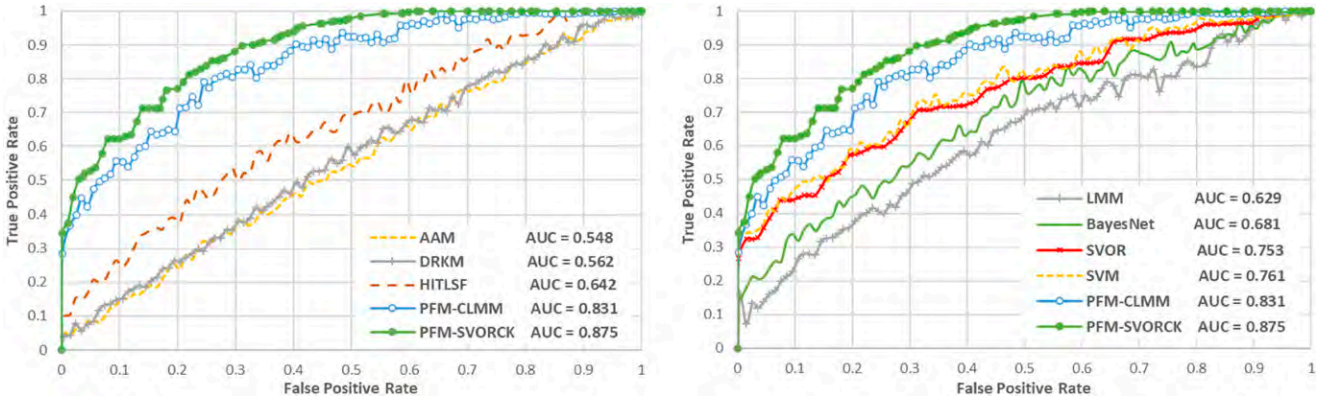
We analyzed the detection performance of PFM (using SVORCK and CLMM) and the top-performing comparison model (HITLSF) and method (SVM) using the 1,421 intention-to-transact instances that transpired during the nine-month test period. The left bars in Figure 6 depict the number and percentage of correctly classified intend-to-transact instances, with PFM detecting 10%–17% more instances than its best competitors. We also extracted a subset of these instances where some transaction behavior was *observed* via the log files, amounting to 1,165 transactions in which the employee either entered information (e.g., in a form or login text box) or agreed to download files or software to the work machine. We examined these observed transactions to see how many were predicted as intention (i.e., the most severe stage in our funnel). As shown in the right bars in Figure 6, PFM also attained markedly better performance on this subset of observed transactions, with detection rates of 90%–94%. Paired t tests revealed that PFM-SVORCK's performance lifts were significant on both intention and observed transactions (all $p < 0.001$, on $n = 1{,}421$ for intention and $n = 1{,}165$ for observed). Similarly, PFM-CLMM also significantly outperformed SVM and HITLSF (all $p < 0.001$).

Table 5. AUC Values for Prediction ROC Curves and p values for PFM and Comparison Models/Methods

Comparison model	AUC	vs. PFM SVORCK	vs. PFM CLMM	Comparison method	AUC	vs. PFM SVORCK	vs. PFM CLMM
PFM-SVORCK	0.875	—	—	PFM-SVORCK	0.875	—	—
PFM-CLMM	0.831	<0.001***	—	PFM-CLMM	0.831	<0.001***	—
HITLSF	0.642	<0.001***	<0.001***	SVM	0.761	<0.001***	<0.001***
DRKM	0.562	<0.001***	<0.001***	SVOR	0.753	<0.001***	<0.001***
AAM	0.548	<0.001***	<0.001***	CLMM-Equi	0.730	<0.001***	<0.001***
				BayesNet	0.681	<0.001***	<0.001***
				LMM	0.629	<0.001***	<0.001***

***$p < 0.001$.

Figure 4. (Color online) ROC Curves of Funnel Stage Predictions Across Models and Methods

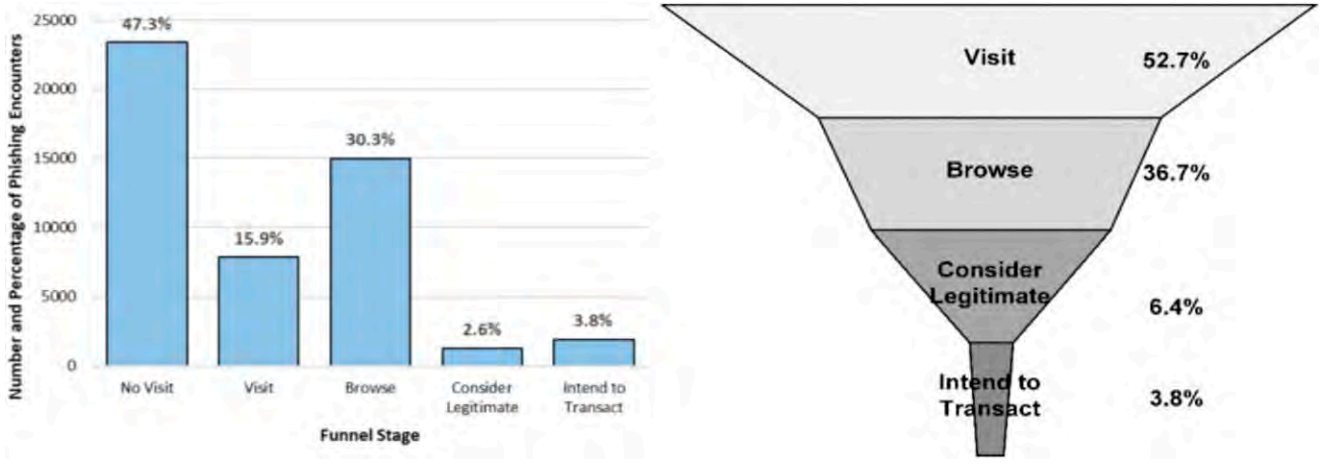


5.2.1. Experiment 1: Prediction—Performance on High-Severity URLs Across Threats and Channels. Regarding visits to high-severity phishing URLs containing malware, Figure 7 depicts the frequency of concocted (Con) and spoof (Spf) sites where PFM, SVM, and HITLSF correctly predicted that the user would at least visit the URL. The bars denote threats encountered via email (work or personal), social media, or search engine results, and threats were also categorized as generic attacks (Gen), spear phishing attacks (SP) tailored toward the organizational context, or watering hole attacks (WH) that use concocted websites. As depicted, PFM outperformed the best comparison model (HITLSF) and method (SVM) on high-severity threats across various communication channels, with the exception of generic spoof attacks appearing in work email. Overall, PFM-SVORCK was able to correctly predict visits to high-severity threats for 96% of the cases in the nine-month test period, which amounts to 170 greater detection occurrences (10% points higher) than the closest competitor. Given the hefty costs exacted by such high-severity threats, these results have important implications for proactive organizational security.

We also examined AUCs within these different threat channels and found that PFM's performance was fairly robust across email, social media, and search engine threats (Table 6). For the four channels, in addition to performance, we report the overall AUC values previously presented in Table 5. Interestingly, both work email and search engine results yielded AUC values that were higher than the overall performance, whereas personal email and social media performed below average, with personal email being the weakest performer (significantly lower). Overall, the lack of significant variation in performance by channels underscores the robustness of PFM's susceptibility prediction capabilities.

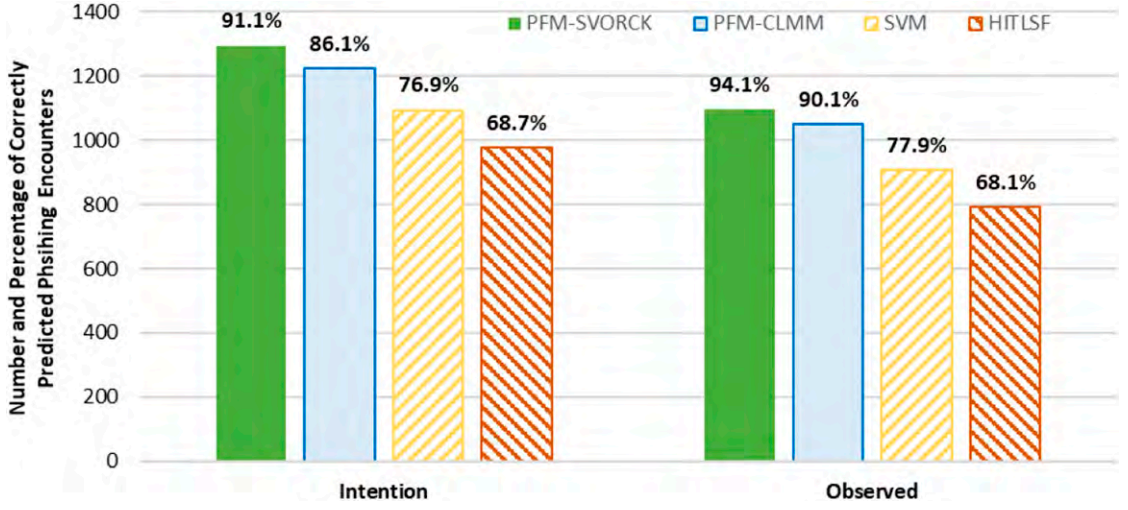
The slightly lower performance on social media and personal email might be explained by the fact that these channels may encompass a more diverse set of threat characteristics and exploitation strategies,

Figure 5. (Color online) Phishing Funnel Stage Traversal Statistics Across 12 Months of Employee-Phish Encounters



Note. Left panel shows quantity of user-phish encounters *ending* at that particular funnel stage; right panel shows funnel with percentages depicting how many sessions went *at least* to that stage.

Figure 6. (Color online) Number and Percentage of Correctly Predicted Employee Intention to Transact (and Observed) Instances

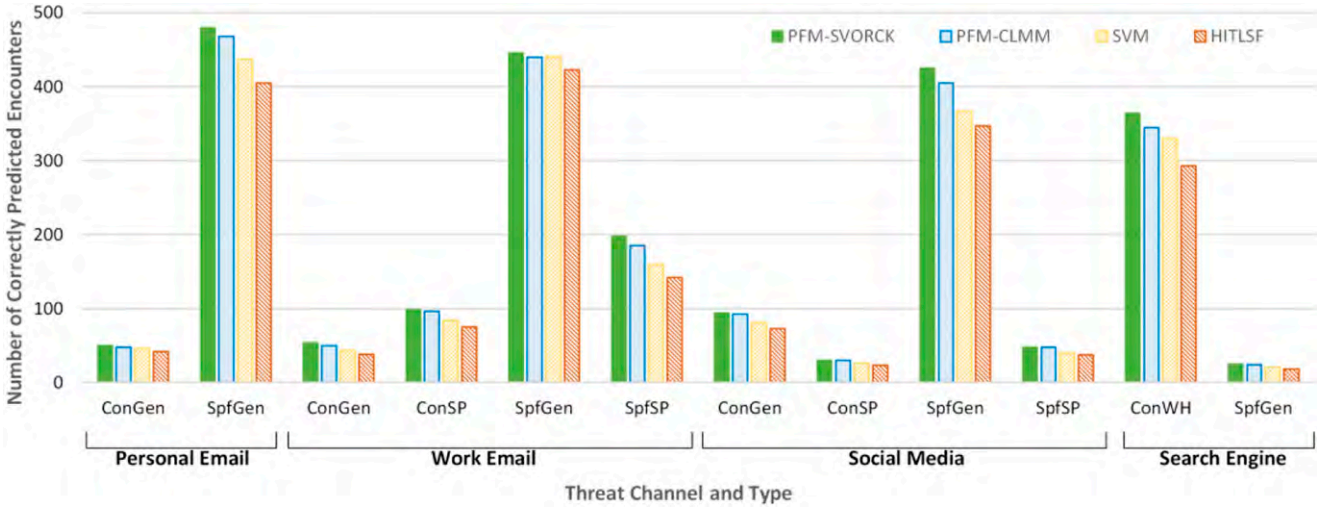


based on personal context factors. Although we measured users' familiarity with many commonly spoofed websites, the email-based phishing literature has mentioned personalized strategies such as social phishing (Jagatic et al. 2007) that might use cues beyond the threat characteristics adopted in our study. Moreover, other research has also examined the role of context with respect to email, such as time of day or number of emails in the inbox (Wang et al. 2012), which may also serve as important cues. Additionally, emails and social media often encompass scams and other visual cues. Scam knowledge and such cues go beyond website familiarity and general phishing awareness (Wang et al. 2012).

It is worth noting that PFM did not explicitly incorporate these channels as a threat characteristic variable—a potential future direction. It is also important to note that our performance regarding email-based attacks might have been enhanced by the fact that PFM only examined emails containing a website URL. There are other email-based attacks involving phone numbers, malicious attachments, and image downloads that are precluded from our field study test beds.

5.2.2. Experiment 1: Prediction—Impact of Features. To examine the utility of the six categories of PFM features for predicting user susceptibility, we examined the performance of PFM using all features versus performance when using all but one category (Table 7). We conducted the evaluation using the exact same longitudinal training and testing setup as

Figure 7. (Color online) Number of Correctly Predicted High-Severity Threats Visited by Employees



Note. Con, concocted; Spf, spoof; SP, spear phishing; Gen, generic attacks; WH, watering hole attacks.

Table 6. AUC Values on Prediction ROC Curves for PFM on Different Threat Channels

PFM method and channels	AUC	vs. all	PFM method and channels	AUC	vs. all
PFM-SVORCK—Search engine	0.903	0.00***	PFM-CLMM—Search engine	0.855	0.00***
PFM-SVORCK—Work email	0.881	0.21	PFM-CLMM—Work email	0.833	0.45
PFM-SVORCK—All channels	0.875	—	PFM-CLMM—All channels	0.831	—
PFM-SVORCK—Social media	0.872	0.29	PFM-CLMM—Social media	0.827	0.20
PFM-SVORCK—Personal email	0.862	0.03*	PFM-CLMM—Personal email	0.822	0.06

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

outlined earlier. The experiment results for PFM-SVORCK and PFM-CLMM are as follows: exclusion of tool performance, tool perception, threat characteristics, prior experiences, and demographics all resulted in significant performance degradation in terms of lower AUC values, both for PFM-SVORCK and PFM-CLMM (all $p < 0.001$). Threat perceptions were also significant ($p = 0.002$) for PFM-SVORCK but not for PFM-CLMM. The results underscore the value of the six feature categories included in PFM. Most categories significantly contributed to the overall susceptibility prediction power of PFM. Moreover, all categories added an AUC lift to overall performance, although in the case of threat perceptions, the lift was not significant for the PFM-CLMM setting.

PFM uses observed and perceptual survey-based variables as input features. To further examine the efficacy of the included survey-based variables, we compared the PFM features against a feature set that also encompassed all the HITLSF, DRKM, and AAM features (see Table C2 in Online Appendix C). This *all variables* feature set included survey-based features for past encounters, risk propensity, security habits, self-efficacy, technical ability, and trust in tool (see Table C1 in Online Appendix C). Because perceptual items entail an additional data collection cost (i.e., surveying employees), we also examined the use of an *observed only* feature set comprising only the 10 observed, nonperceptual features (i.e., those relating to tool performance, threat characteristics, and demographics). Finally, we also supplemented this latter feature set by including data from the five most recent user-phish encounters in a feature set that included the 10 observed features per encounter and the final funnel stage, resulting in 55 total prior log variables. One advantage of reliance on logs is that it may enable faster model update (i.e., retraining on new IVs). Accordingly, rather than retraining every three months, as done with the models using survey variables, we retrained this *observed + prior logs* model every month. All feature sets were run using SVORCK on the longitudinal field data, as done before.

The results comparing these four feature sets appear on the left side of Table 8. Interestingly, the inclusion of the additional survey items in the *all features* setting did not improve performance. Conversely, the AUC was somewhat lower suggesting that some of the additional features developed by competing models may in fact be noisy and less effective for susceptibility prediction. Unsurprisingly, excluding all perceptual features as in the *observed only* setting resulted in a large performance drop—this is consistent with our observations presented in Table 7 when tool perceptions and prior experiences were excluded. Whereas inclusion of prior logs offset this drop to some extent, it was not enough to entirely compensate for the exclusion of perceptual features. These results further underscore the importance of the survey-based features in PFM.

We also explored the impact of feature selection as a means of reducing the feature set (especially the survey-based items). Recursive feature elimination (RFE) was applied using cross-validation within the training data for each window to reduce the feature

Table 7. AUC Values on Prediction ROC Curves for PFM Using Different Feature Categories

PFM method and features	AUC	vs. PFM SVORCK	PFM method and features	AUC	vs. PFM CLMM
PFM-SVORCK	0.875	—	PFM-CLMM	0.831	—
No tool performance	0.816	<0.001***	No tool performance	0.773	<0.001***
No tool perceptions	0.808	<0.001***	No tool perceptions	0.770	<0.001***
No threat characteristics	0.821	<0.001***	No threat characteristics	0.789	<0.001***
No threat perceptions	0.858	0.002**	No threat perceptions	0.821	0.051
No prior experiences	0.810	<0.001***	No prior experiences	0.802	<0.001***
No demographics	0.851	<0.001***	No demographics	0.814	0.004**

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

Table 8. AUC Values on Prediction ROC Curves for SVORCK Using Different Feature Sets

No feature selection	AUC	vs. PFM no feature selection	With feature selection	AUC	vs. PFM with feature selection	vs. PFM no feature selection
All PFM variables	0.875	—	All PFM variables	0.881	—	0.102
All variables	0.860	0.003**	All variables	0.884	0.147	0.079
PFM observed only	0.772	<0.001***	PFM observed only	0.780	<0.001***	<0.001***
PFM observed + prior logs	0.821	<0.001***	PFM observed + prior logs	0.836	<0.001***	<0.001***

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

set (Guyon et al. 2002). We used RFE because it is a multivariate selection method that works well with support vector machines, has yielded good results in prior studies, and attained the best results with our data. The right side of Table 8 shows the results for the four feature sets when using feature selection. The *all variables* setting coupled with feature selection produced the best results, but none of the settings significantly outperformed the PFM variables with or without feature selection (see "vs. PFM no FS" and "vs. PFM with FS" columns for paired *t* test *p* values). The limited lift attributable to the *all variables* with feature selection stemmed from the fact that none of the additional features beyond those appearing in PFM ranked in the top 12 (based on RFE values), with most appearing in the bottom 10.

5.2.3. Experiment 1: Prediction—Robustness of Design. Our field study design entailed quarterly surveys and users were also prompted with a pop-up form after their sessions with potential phishing sites asking them if they considered the site to be legitimate and/or intended to transact with it. These elements of the field study design had the potential to alter employee behavior (e.g., a Hawthorne effect). To examine the potential impact of asking survey questions every three months, we plotted employees' mean monthly funnel traversal behaviors for five possible stages: visit, browse, consider legitimate, intend to transact, and actual (observed) transaction. Figure 8 depicts the results. As shown in the figure, there are no noticeable patterns over the three-month intervals between surveys (i.e., months 1–3, 4–6, 7–9, or 10–12) or across the 12-month time period as a whole. For instance, visitation, browsing, and so on, are not lower in the month immediately following a survey.

Similarly, asking users whether they considered the website to be legitimate or intended to transact with it may have altered their behavior when encountering potential phishing websites. We examined this potential concern by conducting a 3-month pilot study prior to the 12-month longitudinal experiment. A group of 300 employees from FinOrg and LegOrg were invited to participate in the three-month pilot study. These employees did not overlap at all with the ones invited to participate in the subsequent 12-month study and were chosen at random. The pilot study invitees were given the exact same information and incentives as those involved in the full study. A total of 205 employees agreed to participate: they were randomly split into control and treatment groups. During the course of the pilot experiment, three participants left the company for normal attrition reasons. The control group participants did not receive any pop-up forms after their sessions. The treatment group participants did receive the short pop-up forms after each session with a potential phishing website. Figure 9 shows the funnel traversal behavior for the control and treatment groups across all ex post verified user-phish encounters. We observed

Figure 8. (Color online) Mean Monthly Funnel Stage Traversal Probabilities Across 12-Month Field Study

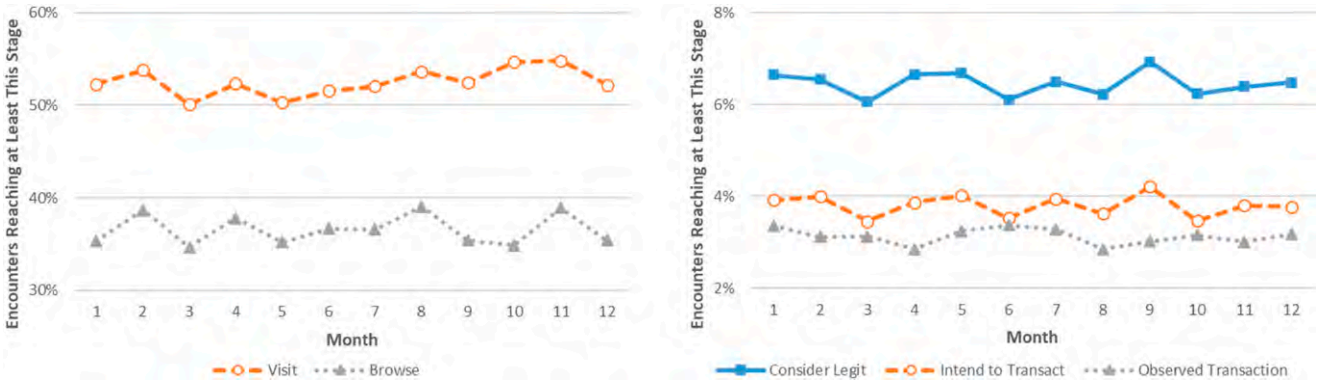
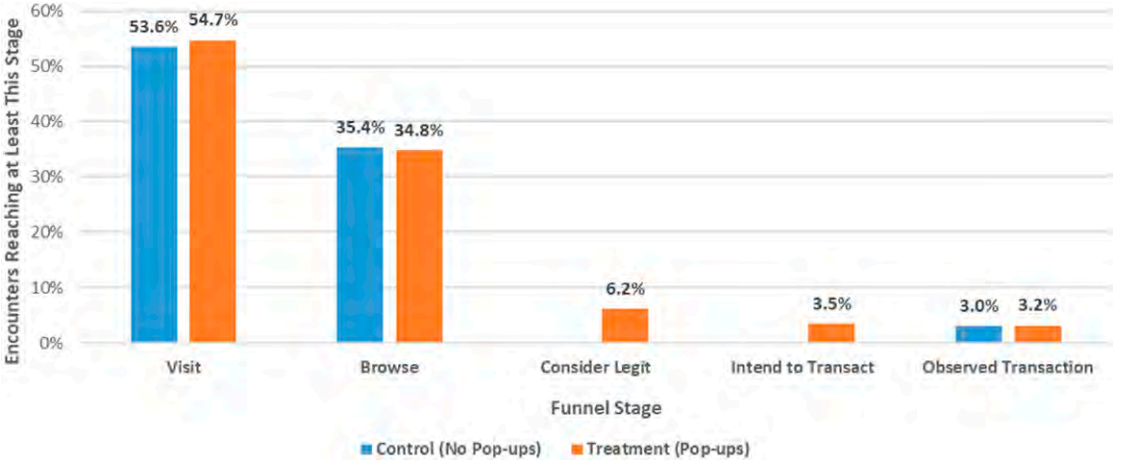


Figure 9. (Color online) Funnel Traversal Behavior for Pilot Study Employees in Control and Treatment Groups



no significant differences between the two groups regarding percentage of employees who visited, browsed, or in observed transactions (i.e., the three decisions not requiring user input). In the absence of the pop-ups, no information was recorded in the control group for the consider legit and intend to transact stages. The pilot results suggest that the postsession pop-up form likely did not alter funnel behavior for those in the treatment group. The observed transactions were also highly correlated with the *intend to transact* and *consider legitimate* values gathered via the pop-up forms for the treatment group. Nevertheless, as with any study leveraging perceptual data, this did not preclude our experiment design from the possibility of certain response biases with respect to the *consider legitimate* and *intend to transact* stages.

6. Experiment 2: Intervention—Field Testing Effectiveness of Prediction-Guided Interventions

Our second research question asked: *How effectively can interventions driven by susceptibility predictions improve avoidance outcomes in organizational settings?* To answer this question related to the downstream value proposition of accurately predicting susceptibility, we followed up our prediction field experiment (described in Section 5) with a longitudinal multivariate field experiment. The field test was performed over a three-month time period at FinOrg and LegOrg using the same set of 1,278 employees incorporated in the prior field experiment. Because of normal workforce attrition and a few opt-out cases, 1,218 employees participated in the experiment. The experiment design and variable operationalizations used were the same as the prior field study. All participants filled out the same survey as prior experiments at the beginning of the three-month period.

6.1. Experiment 2: Intervention—Design

Each participant was randomly assigned to one of six settings for the duration of the experiment: PFM-SVORCK, PFM-CLMM, SVM, HITLSF, random, and standard. Employees in the standard setting represented the status quo control group: these individuals received the default warning for each phishing URL, irrespective of their predicted susceptibility levels. Conversely, the PFM-SVORCK, PFM-CLMM, SVM, and HITLSF groups received one of three warnings (default, medium severity, and high severity) based on their respective model's predicted susceptibility level along the phishing funnel. Aligning warnings with user or other contextual factors has been found to be a potentially effective security intervention, provided that warning fatigue can be properly managed (Chen et al. 2011; Vance et al. 2015, 2018). These warnings differed in terms of size, colors, icons, and message text.

For user-phish encounters predicted to end without a visit, the default warning was displayed. For those predicted to result in visitation and/or browsing, the medium-severity warning was presented. Finally, user-phish encounters predicted to culminate with consider legitimate or intend to transact garnered a high-severity warning. To control for behavioral changes attributable to introduction of the new medium- and high-severity warnings, relative to the default one used in the standard setting, we incorporated an additional random setting. Participants assigned to this setting randomly received either the default, medium-severity, or high-severity warning. Their likelihood of receiving default, medium-severity, and high-severity warnings was based on the overall phishing funnel observed across the 12-month field study (depicted earlier in Figure 5). In other words, for users in this setting, the probability

of receiving a default warning was 47.3%, medium-severity was 46.3%, and high-severity was 6.4%.

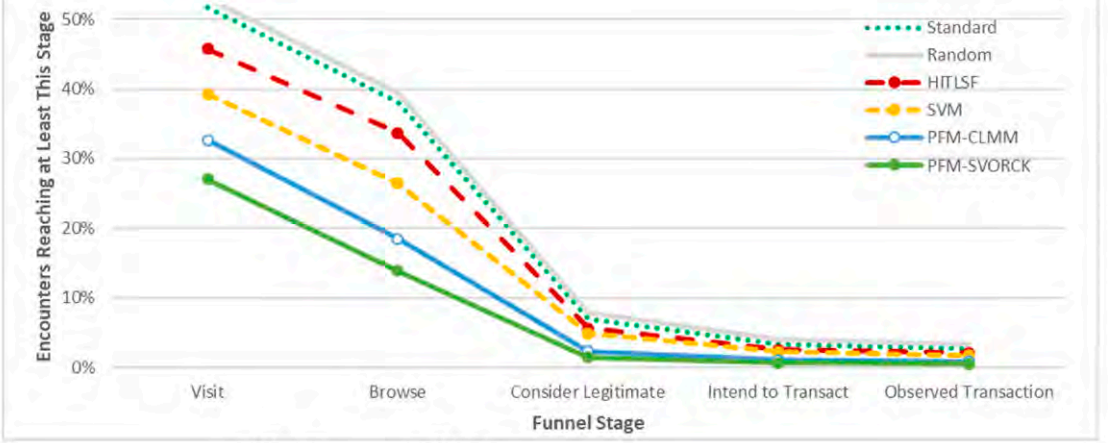
For those employees assigned to the PFM-SVORCK, PFM-CLMM, SVM, and HITLSF settings, data from months 10–12 of the prior experiment was used to train their respective susceptibility prediction model. To reiterate, model predictions were not used for employees in the random and standard settings. During the three-month study, there were an average of 11.35 actual phishing encounters per employee. Phishing emails were verified as described in Section 5.1.

6.2. Experiment 2: Intervention—Results

We evaluated performance by examining actual phishing funnels for participants assigned to the six settings. Figure 10 shows the experiment results depicting the percentage of user-phishing encounters for each of the six settings that went at least as far as that particular funnel stage. Participants using PFM for susceptibility prediction were less likely to traverse the phishing funnel stages and had lower visitation, browsing, legitimacy consideration, and transaction intention rates. On average, PFM outperformed SVM, HITLSF, and the standard setting by 7–20 percentage points at the higher funnel stages and generated less than half the number of traversals for the latter stages of the funnel. The users assigned to the benchmark or baseline settings had three to six times as many observed transactions with phishing websites across the three-month duration of the study, relative to users assigned to PFM-SVORCK. Compared with PFM-CLMM, PFM-SVORCK resulted in 20%–30% fewer visits and browses and 40% fewer transaction intentions and observed transactions. These results highlight the sensitivity of intervention effectiveness to the performance of the underlying predictive models' accuracy in field settings, thereby underscoring the importance of enhanced prediction. Interestingly, the random setting underperformed in comparison with the standard setting, suggesting that displaying alternative warnings without aligning them with predicted susceptibility levels did not improve threat avoidance performance.

To examine the statistical significance of the results presented in Figure 10, we conducted a series of one-way analyses of variance (ANOVAs), comparing outcomes across the six settings at each funnel stage. Based on these ANOVAs, the settings were significantly different at each step of the funnel: visit, $\chi^2(5) = 699.7$, $p < 0.001$; browse, $\chi^2(5) = 800.6$, $p < 0.001$; consider legitimate, $\chi^2(5) = 214.5$, $p < 0.001$; intend to transact, $\chi^2(5) = 101.7$, $p < 0.001$; and observed transaction, $\chi^2(5) = 85.3$, $p < 0.001$. To follow up on these omnibus tests, we conducted two additional sets of contrasts to evaluate the effectiveness of PFM relative to the other settings. First, we compared the average of the two PFM settings (i.e., PFM-SVORCK and PFM-CLMM) to the non-PFM competitor settings

Figure 10. (Color online) Phishing Funnel Traversal Percentages for Employees Assigned to Six Experimental Settings



Funnel Stage	PFM-SVORCK	PFM-CLMM	SVM	HITLSF	Random	Standard
Visit	27.02	32.63	39.27	45.78	53.08	51.76
Browse	13.88	18.44	26.46	33.72	39.37	38.05
Consider Legitimate	1.45	2.28	4.83	5.59	7.84	6.92
Intend to Transact	0.61	1.07	2.24	2.50	4.03	3.25
Observed Transaction	0.50	0.86	1.74	2.10	3.27	2.64

Note. The chart/table depict the percentage of all user-phish encounters that went *at least* to that stage of the funnel.

(i.e., SVM, HITLSF, random, and standard). Each of these comparisons was significant at every funnel stage using Bonferroni adjusted p values: visit, $\chi^2(1) = 200.4$, $p < 0.001$; browse, $\chi^2(1) = 234.3$, $p < 0.001$; consider legitimate, $\chi^2(1) = 68.4$, $p < 0.001$; intend to transact, $\chi^2(1) = 32.1$, $p < 0.001$; and observed transaction, $\chi^2(1) = 26.2$, $p < 0.001$. Second, we compared PFM-SVORCK versus PFM-CLMM directly to determine which setting performed best overall. In these comparisons, PFM-SVORCK outperformed PFM-CLMM in all funnel stages except observed transaction: visit, $\chi^2(1) = 35.6$, $p < 0.001$; browse, $\chi^2(1) = 28.3$, $p < 0.001$; consider legitimate, $\chi^2(1) = 7.4$, $p = 0.007$; intend to transact, $\chi^2(1) = 4.9$, $p = 0.027$; and observed transaction, $\chi^2(1) = 2.9$, $p = 0.090$.

Collectively, these contrasts showed (1) that PFM settings outperformed competitor settings and (2) that PFM-SVORCK significantly enhanced susceptibility avoidance performance over PFM-CLMM for the visit, browse, consider legitimate, and intention to transact stages.

6.2.1. Experiment 2: Intervention—Cost-Benefit of Interventions Guided by Susceptibility Predictions. Prior design science studies have shown that cost-benefit analysis is useful for examining the practical value of design artifacts deployed in the field (Kitchens et al. 2018). In the case of predicting phishing susceptibility, monetary benefits can be quantified as the savings attributable to reduced funnel traversal behavior (Canfield and Fischhoff 2018). Each time a user avoids the funnel stages of visiting, browsing, or transacting with a phishing site, there is a cost-savings benefit to the firm.

For example, FinOrg estimated that, on average, each avoided employee visit to a verified phishing website saved HelpDesk/tech support one hour of time and effort (about $70). This time and effort savings increased to 1.5 hours for instances in which the user would have browsed on the site. Further, using FinOrg's conservative estimate, avoiding a single observed user transaction resulted in a median of $1,000 in savings on security patching and remediation.[1] The total estimated annual phishing-related costs at FinOrg were $32 million, compared with an estimated $25 million average annual cost of phishing for U.S.-based financial services firms (Richards et al. 2017).

However, unnecessary interventions resulting from overestimated susceptibility predictions (i.e., predicting users to go further down the funnel than they actually would have) can also lead to interruptions, productivity losses, and unnecessary labor costs (Jenkins et al. 2016, Richards et al. 2017). FinOrg believed that displaying a higher-severity warning unnecessarily (i.e., medium or high when the actual susceptibility level was low) reduced productivity by one hour because of employee interruptions, seeking HelpDesk support, clarifications, and so on (Canfield and Fischhoff 2018). Each such user-phish incident cost the firm an estimated $50.

We examined the monetary benefit to firms such as FinOrg/LegOrg of aligning interventions (in our case, warning severity) with user susceptibility levels. We projected the results of our three-month field intervention study to annual monetary business value for FinOrg, a large firm with 10,000 corporate employees that routinely uses company-issued desktop/laptop devices for work. For our cost-benefit analysis, the status quo was the standard setting in which employees used the existing enterprise security solution featuring the default warning. We evaluated the monetary value of the other five settings (i.e., PFM-SVORCK, PFM-CLMM, SVM, HITLSF, and random) relative to the standard setting. Specifically, we calculated funnel avoidance benefits as reductions in visitation, browsing, and observed transactions with verified phishing websites for the five treatment settings, relative to the standard setting.

Table 9 shows the estimated annual benefits for the five treatment settings. Based on less visitation, browsing, and transactions with phishing websites, use of PFM-SVORCK could yield $1,960 in benefits per employee. Conversely, because of a large number of false high-severity warnings (SVM) and medium-severity

Table 9. Estimated Annual Benefit of Interventions Driven by Susceptibility Predictions for FinOrg

	PFM-SVORCK	PFM-CLMM	SVM	HITLSF	Random
Benefits per employee					
Fewer visits	$673	$520	$345	$165	($27)
Less browsing	$329	$267	$159	$60	($15)
Fewer observed transactions	$1,163	$966	$493	$296	($335)
Costs per employee					
Unnecessary severe warnings	($204)	($298)	($928)	($717)	($908)
Gross annual benefit					
Per employee	$1,960	$1,454	$68	($198)	($1,284)
FinOrg total (10K employees)	$19,603,941	$14,542,857	$682,759	($1,975,369)	($12,839,409)

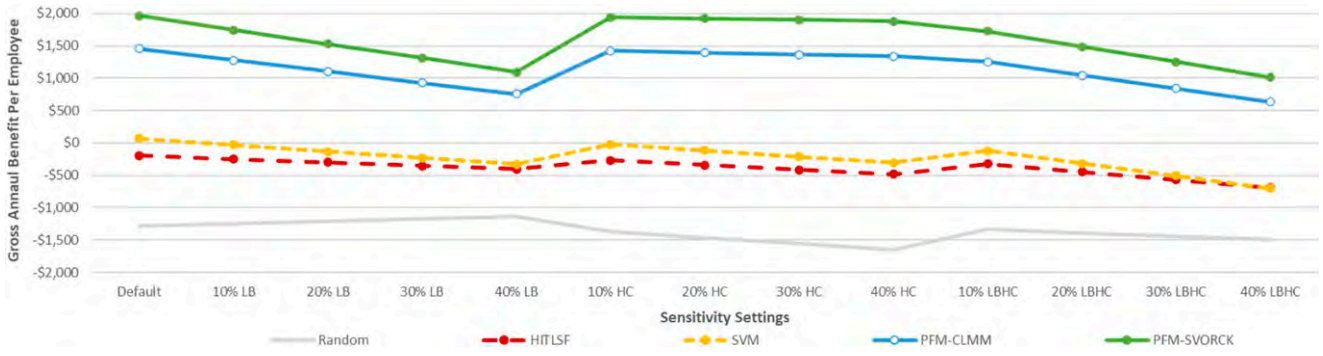
warnings (HITLSF), employees assigned to these methods may suffer exceedingly high levels of warning fatigue and false positives. In the case of HITLSF, these costs outweighed the avoidance benefits. The random setting quantified the cost of arbitrarily displaying higher severity warnings. Relative to PFM-CLMM, the PFM-SVORCK setting garnered a lift of $500 per employee—an additional potential benefit of $5 million annually for FinOrg.

To examine the sensitivity of gross annual benefits per employee, we assessed the impact of lower benefits (LB), higher costs (HC), or both (i.e., lower benefits and higher costs (LBHC)). For the LB setting, we held the costs of unnecessary warnings constant but assumed that fewer visits, browsing, and observed transaction-related benefits would be 10%40% lower in 10 percentage point intervals. Similarly, in the HC setting, we increased the cost of unnecessary warnings by 10%, 20%, 30%, or 40% while holding the benefits constant at the default level. For the LBHC setting, we both reduced benefits and increased costs by x% at the same time. The results for these 12 settings and the default cost-benefit assumption levels depicted earlier in Table 9 all appear in Figure 11. As shown in the figure, even when looking at an extreme scenario such as reducing the potential benefits of intervention by 40% while simultaneously increasing the costs of unnecessary warnings by 40%, PFM-SVORCK still provides a gross annual benefit per employee of more than $1,000 (PFM-CLMM provides a benefit of $634), whereas comparison methods such as SVM and HITLSF generate losses of around $700 per employee. The results suggest that the gains associated with PFM are fairly resilient across a wide range of cost-benefit values.

This analysis is not without caveats. First, because of differences in firm size and industry sectors, the annual benefit for other organizations may vary. For instance, although the estimated per employee differences at LegOrg are slightly higher in favor of PFM-SVORCK, the annual benefit relative to PFM-CLMM and SVM is $3 million and $10 million, respectively (not reported in Table 9). Second, the analysis focuses on gross benefit, whereas the cost of implementing any susceptibility prediction solution—along with training employees and embedding a user response team—can cost $200–$300 per employee annually. Nevertheless, the results clearly illustrate the benefit of accurately predicting phishing susceptibility and suggest that this type of approach can be a valuable component of an enterprise anti-phishing strategy.

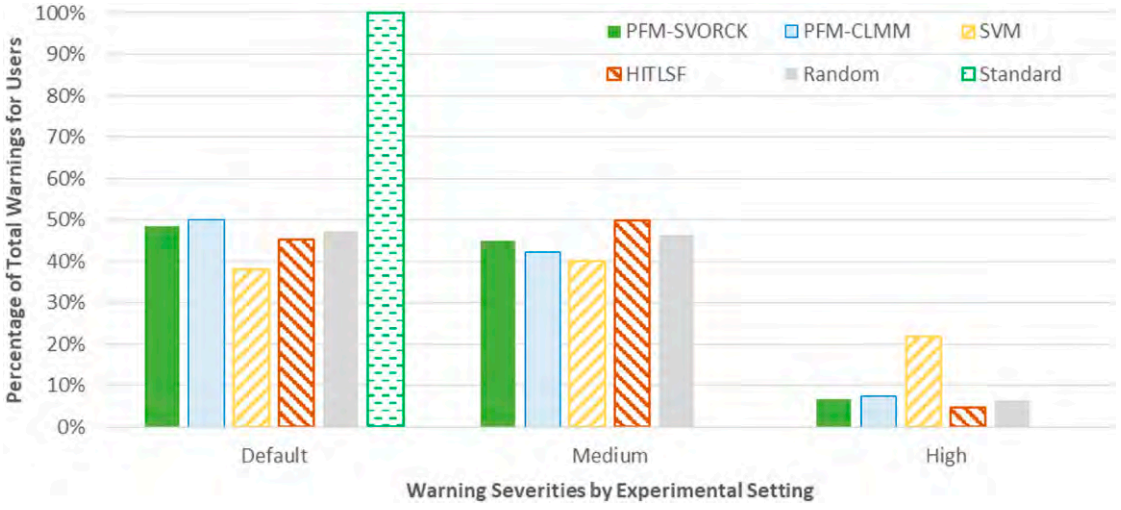
6.2.2. Experiment 2: Intervention—Robustness of Design. To ensure that the results attained in Figure 10 were not simply attributed to the quantity of default, medium-severity, and high-severity warnings seen by employees assigned to the six experimental settings versus the alignment between user susceptibility to that particular threat and warning severity, we examined the percentage of types of warnings displayed to the six groups. Because the *total* number of warnings was not significantly different across the six settings, for ease of interpretation, percentages were used, as opposed to raw counts. Figure 12 displays the results. As noted, users in the random setting randomly received the default, medium-severity, and high-severity warnings proportionally to the funnel traversal behaviors in the 12-month prediction study. With respect to high-severity warnings, users assigned to the SVM setting received the most, whereas those in the HITLSF setting received the least (with the exception of those assigned to the standard setting control group—that group saw only the default warning throughout). Relative to those in the SVM, HITLSF, and random settings, users in the PFM-SVORCK and PFM-CLMM settings received the highest proportion of default warnings.

Figure 11. (Color online) Sensitivity of Gross Annual Benefit Per Employee to Cost-Benefit Assumptions



Note. LB, lower benefit; HC, higher costs; LBHC, lower benefit and higher costs.

Figure 12. (Color online) Percentage of Default, Medium-Severity, and High-Severity Warnings Displayed to Employees Assigned to Six Experimental Settings



These results suggest that the avoidance behaviors observed in the prior section (Figure 9) for warnings guided by PFM were not attributable to the quantity of medium- or high-severity warnings displayed.

7. Results Discussion and Concluding Remarks

7.1. Results Summary

Our experiments demonstrate the utility of PFM, which incorporates tool, threat, and user-related variables to predict phishing funnel stages for user-phish encounters. Managers tasked with enterprise security recognize the need for a multipronged approach encompassing the adoption of appropriate security IT artifacts, policies/procedures, and compliance/protective behavior (Ransbotham and Mitra 2009, Santhanam et al. 2010, Wright et al. 2014). Table 10 summarizes our key findings.

For RQ1, experiment 1, our 12-month longitudinal field experiment showed that PFM significantly outperforms competing models in predicting employees' phishing susceptibility in organizational settings, thus reinforcing PFM's potential for offering real-time, preventative solutions based on its predictive merits. Specifically, PFM obtained an AUC score that was 8%–52% higher than those of competing models/methods, correctly predicting visits to high-severity threats for 96% of the cases over the nine-month test period—a result that was 10% points higher than the nearest competitor. The windowing approach used for model training/testing also lends credence to PFM's potential to adapt to changes in user behavior or the environment that occur over time.

Table 10. Summary of Key Findings Pertaining to Proposed PFM

Research question	Key results
RQ1: How effectively can PFM predict users' phishing susceptibility over time and in organizational settings?	1. Over a nine-month test period, PFM outperformed competing models in predicting employees' phishing susceptibility at two organizations. 2. PFM's AUC scores were 8%–52% higher than competing models, and PFM correctly predicted visits to high-severity threats for 96% of cases—a result 10 percentage points higher than the best comparison method. 3. PFM performed better on an array of threats across search, social, web, and email-based attacks. 4. Feature impact analysis showed that all categories of features in PFM significantly contributed to overall predictive power.
RQ2: How effectively can interventions driven by susceptibility predictions improve avoidance outcomes in organizational settings?	1. Over a three-month period, participants using PFM-SVORCK for susceptibility prediction were significantly less likely to traverse the phishing funnel stages, with lower visitation, browsing, legitimacy consideration, transaction intention, and observed transaction rates. 2. Cost-benefit analysis revealed that interventions guided by PFM-SVORCK resulted in gross annual phishing-related cybersecurity cost reductions of nearly $1,900 per employee more than comparison prediction methods, and $500 more than the PFM-CLMM setting.

For RQ2, experiment 2, our three-month longitudinal field experiment showed the efficacy of interventions guided by accurate and personalized real-time susceptibility prediction. Previous research has suggested that users ignore anti-phishing tool warnings because they are not personalized to themselves (Chen et al. 2011). In contrast, participants using PFM for susceptibility prediction viewed warnings that were more congruent with their susceptibility to the impending threat and they were consequently less likely to traverse the phishing funnel stages, resulting in lower visitation, browsing, legitimacy consideration, transaction intention, and observed transaction rates. Users equipped with PFM-driven warnings were one half to one third as likely to transact with phishing threats, thereby demonstrating the downstream value proposition of effective and personalized real-time susceptibility prediction.

These results open up possibilities not only for proactive identification of susceptible users but also for a bigger-picture approach involving personalized real-time security warnings and/or access control policies based on predicted susceptibility in organizational settings. For example, given PFM's capacity to perform real-time prediction, an organization's IT security policy might entail temporarily—but, more importantly, immediately—blocking user access when an employee is traversing deeper into a social media phishing funnel threat to avoid the most dangerous outcomes. Such a policy would also include sterner warnings and/or escalating access restrictions for negligent or otherwise intransigent users who are predicted to be at greatest risk of a future security breach. In fact, equipped with robust predictive capabilities, FinOrg and LegOrg are currently exploring these types of real-time protective measures.

7.2. Contributions

In this study, we proposed PFM as a design artifact for predicting user susceptibility to phishing website-based attacks. The major contributions of our work are threefold. First, given the need for mechanisms capable of modeling behavior in relevant security contexts (Wang et al. 2015), we developed the PFM design artifact, which incorporates the phishing funnel as a mechanism for representing users' key decisions and actions when encountering phishing websites. PFM uses theoretically motivated set of decision/action predictors including tool, threat, and user-related attributes. We estimated PFM using a novel support vector ordinal regression with a composite kernel (SVORCK) capable of parsimoniously considering user-phishing interactions and funnel stage traversal behaviors.

Second, to evaluate the modeling and prediction merits of PFM, we performed two large-scale, longitudinal field experiments. Experiment 1 comprised a longitudinal field experiment conducted over the course of 12 months in two different organizations involving 1,278 employees and 49,373 phishing interactions. PFM substantially outperformed competing models in terms of predicting both phishing susceptibility intention and behavior. Experiment 2 involved a second three-month field study in the same two organizations using 1,218 employees and 13,824 user-phish encounters. Warnings guided by PFM's predictions resulted in markedly enhanced threat avoidance behavior resulting in lower visitation, browsing, legitimacy consideration, intention to transact, and observed transactions.

The development of PFM follows guidelines mentioned in recent design science papers that promote the development of novel design artifacts (Gregor and Hevner 2013, Goes 2014). Based on these guidelines, PFM's enhanced phishing susceptibility model performance represents an *improvement* contribution. Whereas susceptibility to phishing is a well-known problem, methods geared toward predicting susceptibility and using those predictions for personalized real-time interventions represent a new solution. Our work also follows the IS community guidelines for predictive analytics research (Shmueli and Koppius 2011), a relatively underexplored but increasingly important research area (Abbasi et al. 2015).

Third, we also make several contributions to the online security domain. The predictive possibilities afforded by PFM have important implications for various practitioner groups, particularly in light of the recent industry trend toward security analytics (Chen et al. 2012, Musthaler 2013, Taylor 2014). Phishing attacks impact at least four types of organizations. They affect user trust in (1) security software companies such as McAfee and Symantec and (2) browser developers such as Microsoft and Google (Akhawe and Felt 2013). Phishing also tarnishes the brand equity and customer satisfaction of (3) spoofed companies, such as eBay and JP Morgan Chase (Hong 2012, Shields 2015). When employees access phishing sites from work, they risk compromising (4) their own organization's security.

Given the effectiveness of PFM, an obvious question is why not automatically remove suspected phishing emails and not involve users at all in this decision. As Anderson et al. (2016b, p. 3) note, "Security systems would ideally detect and prevent a threat without user intervention. However, many situations require the user to make a security judgment based on contextual factors." Phishing is one such situation because "a human may be a better judge than a computer about whether an email attachment is suspicious in a particular context" (Cranor 2008, p. 1). Because of the highly contextual nature of phishing, false positives

are inevitable for any phishing detection system. In such cases, if users are not given the option of viewing emails or sites they are sure are legitimate, they are likely to switch to a less restrictive web browser or email client (Felt et al. 2015). In enterprise settings, this may lead to employee dissatisfaction (Kirlappos et al. 2013) or unsecure workarounds (Sarkar et al. 2020).

Nonetheless, our findings could be used in several ways to further future employee and/or customer-facing anti-phishing strategies, including implementing personalized real-time warnings, access controls, and data security policies that adapt over time. For example, selectively blocking access in situations where anti-phishing tool confidence is high and susceptibility predictions are also severe might be a worthwhile future endeavor to consider. This is analogous to the *prioritizing advice* concept that prior work has advocated as a way of aligning organizational security concerns with employee bandwidth constraints (Herley 2009, p. 143). Susceptibility prediction provides an additional tool that can be used to balance phishing-related sociotechnical tensions with compliance and productivity.

7.3. Limitations and Future Work

Our work is not without its limitations. The phishing funnel presently concludes at intention to transact. Research has shown that there is an intention-behavior gap that can manifest in unpredictable ways (Crossler et al. 2014). In our field experiment settings, those intending to transact did not always actually do so (15%–20% did not). However, we believe this issue was partly mitigated by the fact that by accurately predicting funnel traversal behavior all the way to intention to transact, PFM also performed better on user-phish encounters resulting in observed transactions (Figure 8). Moreover, our customized warning interventions were also able to reduce transaction behavior (Figure 9). Nevertheless, future work that formally includes transaction behavior as a funnel stage in the model would allow for a more holistic representation of decision-stages related to susceptibility.

Additionally, PFM was examined in two field settings featuring employees of firms in the financial services and legal industries. Future work is needed to examine the generalizability of PFM to other contexts (e.g., leisure surfing) and target populations (e.g., different types of Internet users). Our field study necessitated periodic surveys and occasional pop-up questions, which may have affected employee behavior. We attempted to mitigate this concern by conducting multiple field studies that built on each other over a 15-month period. We also analyzed funnel traversal behavior over 12 months and did not observe any effects related to the quarterly surveys or over time (see Section 5.2.3). Furthermore, a pilot field study showed that use of pop-up forms did not significantly alter the observed stages of *visit*, *browse*, and *observed transactions*. However, future field studies might be needed to explore behavior effects of susceptibility prediction that entail primary versus secondary data, including the potential for response bias in self-reporting on the *consider legitimate* and *intend to transact* stages.

Future work should consider the tradeoffs in predictive power relative to survey collection lag time and model retraining rate. Feature subset selection may be a worthwhile future direction as well. Section 5.2.2 shows that subset selection can further enhance AUC values by removing noisy survey variables, thereby potentially enhancing prediction and shortening survey lengths. Furthermore, our implementation of comparison susceptibility models involved some adaptations based on differences in context, as noted in Online Appendix C. Additionally, although our cost-benefit analysis presented in Section 6.2.1 demonstrated that PFM-SVORCK is capable of generating significant savings, future work should focus on making costs a core part of the model training process (Abbasi et al. 2012b, Fang 2012). Finally, in the intervention field study, we connected susceptibility predictions to warnings as a whole (Desolda et al. 2019)—future work could explore the interplay between predictions and warning severity at the design element level comprising text, icons, and so on (Chen et al. 2011). Despite these limitations, in response to calls for studies that use field data to better understand employee security (Mahmood et al. 2010, Wang et al. 2015) and the need for security analytics research (Taylor 2014), we believe that the current study constitutes an important first step toward improving predictions of user susceptibility to phishing—a problem that continues to exact significant monetary and social costs.

Acknowledgments

The authors thank Yan Chen and the SE, AE, and anonymous reviewers for their invaluable feedback across multiple versions of the manuscript.

Endnote

[1] The $1,000 was calculated as FinOrg's estimate of 2.86% of observed transactions resulting in a breach × $35,071, the median cost of a breach at FinOrg. We say *conservative* because we used the median instead of the mean; FinOrg observed a long tail with some incidents having a much higher cost. These numbers are consistent with practitioner research. A 2016 Verizon report estimates that 2.2% of observed transactions lead to a breach, and another report by the Ponemon Institute and Accenture estimated the average cost of a phishing breach to be $105,900 (Richards et al. 2017). Hence, transacting with a phish could cost $2,329 on average.

References

Abbasi A, Lau RY, Brown DE (2015) Predicting behavior. *IEEE Intelligence Systems* 30(3):35–43.

Abbasi A, Zahedi FM, Chen Y (2012a) Impact of anti-phishing tool performance on attack success rates. *Proc. IEEE Internat. Conf. on Intelligence and Security Informatics* (IEEE, Piscataway, NJ), 12–17.

Abbasi A, Albrecht C, Vance A, Hansen J (2012b) Metafraud: A meta-learning framework for detecting financial fraud. *Management Inform. Systems Quart.* 36(4):1293–1327.

Abbasi A, Zhang Z, Zimbra D, Chen H, Nunamaker JF Jr (2010) Detecting fake websites: The contribution of statistical learning theory. *Management Inform. Systems Quart.* 34(3):435–461.

Agarwal A, Hosanagar K, Smith MD (2011) Location, location, location: An analysis of profitability of position in online advertising markets. *J. Marketing Res.* 48(6):1057–1073.

Akhawe D, Felt AP (2013) Alice in warningland: A large-scale field study of browser security warning effectiveness. *Proc. 22nd USENIX Security Sympos.* (USENIX Association, Berkeley, CA).

Alnajim A, Munro M (2009) Effects of technical abilities and phishing knowledge on phishing websites detection. *Proc. IASTED Internat. Conf on Software Engineering* (ACTA Press, Calgary, AB, Canada), 120–125.

Anderson B, Vance A, Kirwan B, Eargle D, Jenkins J (2016a) How users perceive and respond to security messages: A NeuroIS research agenda and empirical study. *Eur. J. Inform. Systems* 25(4):364–390.

Anderson BB, Jenkins JL, Vance A, Kirwan CB, Eargle D (2016b) Your memory is working against you: How eye tracking and memory explain habituation to security warnings. *Decision Support Systems* 92(0):3–13.

Angst CM, Agarwal R (2009) Adoption of electronic health records in the presence of privacy concerns: The elaboration likelihood model and individual persuasion. *Management Inform. Systems Quart.* 33(2):339–370.

Bansal G, Zahedi FM, Gefen D (2010) The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. *Decision Support Systems* 49(2):138–150.

Bar-Ilan J, Keenoy K, Levene M, Yaari E (2009) Presentation bias is significant in determining user preference for search results A user study. *J. Amer. Soc. Inform. Sci. Tech.* 60(1):135–149.

Bardhan I, Oh JH, Zheng Z, Kirksey K (2015) Predictive analytics for readmission of patients with congestive heart failure. *Inform. Systems Res.* 26(1):19–39.

Benbasat I, Barki H (2007) Quo vadis TAM? *J. Assoc. Inform. Systems* 8(4):7.

Bishop M, Engle S, Peisert S, Whalen S, Gates C (2009). Case studies of an insider framework. *Proc. 42nd Hawaii Internat. Conf. on System Sciences* (IEEE, New York), 1–10.

Boss S, Galletta D, Lowry PB, Moody GD, Polak P (2015) What do systems users have to fear? Using fear appeals to engender threats and fear that motivate protective security behaviors. *Management Inform. Systems Quart.* 39(4):837–864.

Bravo-Lillo C, Cranor LF, Downs JS, Komanduri S (2011) Bridging the gap in computer security warnings: A mental model approach. *IEEE Security Privacy* 9(2):18–26.

Burges CJ (1998) A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* 2(2):121–167.

Camp LJ (2009) Mental models of privacy and security. *IEEE Tech. Soc. Magazine* 28(3):37–46.

Canfield CI, Fischhoff B (2018) Setting priorities in behavioral interventions: An application to reducing phishing risk. *Risk Analysis* 38(4):826–838.

Cavusoglu H, Mishra B, Raghunathan S (2005) The value of intrusion detection systems in information technology security architecture. *Inform. Systems Res.* 16(1):28–46.

Chen H, Chiang RH, Storey VC (2012) Business intelligence and analytics: From big data to big impact. *Management Inform. Systems Quart.* 36(4):1165–1188.

Chen Y, Zahedi FM, Abbasi A (2011) Interface design elements for anti-phishing systems. *Internat. Conf. on Design Science Research in Information Systems* (Springer, Berlin), 253–265.

Christensen RHB (2015) Analysis of ordinal data with cumulative link models—Estimation with the R-package ordinal. https://mran.microsoft.com/snapshot/2017-12-11/web/packages/ordinal/vignettes/clm_intro.pdf.

Chu W, Keerthi SS (2007) Support vector ordinal regression. *Neural Comput.* 19(3):792–815.

Chua CEH, Wareham J (2004) Fighting internet auction fraud: An assessment and proposal. *IEEE Comput.* 37(10):31–37.

Cram WA, D'arcy J, Proudfoot JG (2019) Seeing the forest and the trees: A meta-analysis of the antecedents to information security policy compliance. *Management Inform. Systems Quart.* 43(2):525–554.

Cranor L (2008). A framework for reasoning about the Human in the Loop. *Proc. 1st Conf. on Usability, Psychology, and Security* (USENIX Association, Berkeley, CA).

Crossler RE, Long JH, Loraas TM, Trinkle BS (2014) Understanding compliance with bring your own device policies utilizing protection motivation theory: Bridging the intention-behavior gap. *J. Inform. Systems* 28(1):209–226.

Cummings A, Lewellen T, McIntire D, Moore A, Trzeciak R (2012) Insider threat study: Illicit cyber activity involving fraud in the U.S. financial services sector. Report, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Management Inform. Systems Quart.* 13(3):319–340.

Desolda G, Di Nocera F, Ferro L, Lanzilotti R, Maggi P, Marrella A (2019) Alerting users about phishing attacks. *Internat. Conf. on Human-Computer Interaction* (Springer, Cham, Switzerland), 134–148.

Dhamija R, Tygar JD, Hearst M (2006) Why phishing works. *Proc. SIGCHI Conf. on Human Factors in Computing Systems* (ACM, New York), 581–590.

Dinev T (2006) Why spoofing is serious Internet fraud. *Commun. ACM* 49(10):76–82.

Downs JS, Holbrook MB, Cranor LF (2006). Decision strategies and susceptibility to phishing. *Proc. Sympos. on Usable Privacy and Security* (USENIX Association, Berkeley, CA), 79–90.

Downs JS, Holbrook M, Cranor LF (2007). Behavioral response to phishing risk. *Proc. ACM Anti-Phishing Working Groups Annu. eCrime Researchers Summit* (ACM, New York), 37–44.

Egelman S, Cranor LF, Hong J (2008). You've been warned: An empirical study of the effectiveness of web browser phishing warnings. *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems* (ACM, New York), 1065–1074.

Fang X (2012) Inference-based naive Bayes: Turning naive Bayes cost-sensitive. *IEEE Trans. Knowledge. Data Engrg.* 25(10):2302–2313.

Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.* 27(8):861–874.

Felt AP, Ainslie A, Reeder RW, Consolvo S, Thyagaraja S, Bettes A, Harris H, et al. (2015) Improving SSL warnings. *Proc. ACM Conf. on Human Factors in Computing Systems*, 2893–2902.

Floyd D, Prentice-Dunn S, Rogers R (2000) A meta-analysis of research on protection motivation theory. *J. Appl. Soc. Psychol.* 30(2):407–429.

Freed L (2011) Managing Forward: Customer Satisfaction as a Predictive Metric for Banks. *U.S. ForeSee Results 2011 Online Banking Study* (ACM, New York), http://bankblog.optirate.com/wp-content/uploads/2011/07/u.s.-foresee-results-2011-online-banking-study.pdf.

Gartner (2011) Magic quadrant for web fraud detection, April 19, 2011. https://www.gartner.com/en/documents/1641814/magic-quadrant-for-web-fraud-detection.

Gefen D, Straub D (1997) Gender differences in the perception and use of email: An extension to the technology acceptance model. *Management Inform. Systems Quart.* 21(4):389–400.

Goes P (2014) Editor's comments: Design science research in top information systems journals. *Management Inform. Systems Quart.* 38(1):iii–viii.

Grazioli S, Jarvenpaa SL (2000) Perils of Internet fraud: An empirical investigation of deception and trust with experienced Internet consumers. *IEEE Trans. Systems Man Cybernetics Part A* 30(4): 395–410.

Grazioli S, Jarvenpaa SL (2003) Consumer and business deception on the internet: Content analysis of documentary evidence. *Internat. J. Electron. Commerce* 7(4):93–118.

Gregor S, Hevner AR (2013) Positioning and presenting design science research for maximum impact. *Management Inform. Systems Quart.* 37(2):337–355.

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. *Machine Learning* 46(1-3):389–422.

Gyongyi Z, Garcia-Molina H (2005) Spam: It's not for inboxes anymore. *IEEE Comput.* 38(10):28–34.

Herath T, Chen R, Wang J, Banjara K, Wilbur J, Rao HR (2014) Security services as coping mechanisms: An investigation into user intention to adopt an email authentication service. *Inform. Systems J.* 24(1):61–84.

Herley C (2009) So long, and no thanks for the externalities: The rational rejection of security advice by users. *Proc. Workshop on New Security Paradigms*, 133–144.

Herzberg A, Jbara A (2008) Security and identification indicators for browsers against spoofing and phishing attacks. *ACM Trans. Internet Tech.* 8(4):1–36.

Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. *Management Inform. Systems Quart.* 28(1):75–105.

Hong J (2012) The state of phishing attacks. *Commun. ACM* 55(1):74–81.

Jagatic TN, Johnson NA, Jakobsson M, Menczer F (2007) Social phishing. *Commun. ACM* 50(10):94–100.

Jenkins J, Anderson B, Vance A, Kirwan B, Eargle D (2016) More harm than good? How security messages that interrupt make us vulnerable. *Inform. Systems Res.* 27(4):880–896.

Jensen ML, Lowry PB, Burgoon JK, Nunamaker JF (2010) Technology dominance in complex decision making: The case of aided credibility assessment. *J. Management Inform. Systems* 27(1):175–202.

Jobber D, Ellis-Chadwick F (1995) *Principles and Practice of Marketing* (McGraw-Hill, New York).

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–292.

Kaushik A (2011) *Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity* (Wiley Publishing, New York).

Keith M, Shao B, Steinbart P (2009) A behavioral analysis of passphrase design and effectiveness. *J. Assoc. Inform. Systems* 10(2): 63–89.

Kirlappos I, Beautement A, Sasse MA (2013) "Comply or die" is dead: Long live security-aware principal agents. *Internat. Conf. on Financial Cryptography and Data Security* (Springer, Berlin), 70–82.

Kitchens B, Dobolyi D, Li J, Abbasi A (2018) Advanced customer analytics: Strategic value through integration of relationship-oriented big data. *J. Management Inform. Systems* 35(2):540–574.

Kolari P, Finin T, Joshi A (2006) SVMs for the blogosphere: Blog identification and splog detection. *AAAI Spring Sympos.: Computational Approaches to Analyzing Weblogs*, 92–99.

Korolov M (2015). Phishing is a $3.7-million annual cost for average large company. *CSO* (August 26). Accessed October 7, 2018, https://www.csoonline.com/article/2975807/phishing-is-a-37-million-annual-cost-for-average-large-company.html

Kumar N, Mohan K, Holowczak R (2008) Locking the door but leaving the computer vulnerable: Factors inhibiting home users' adoption of software firewalls. *Decision Support Systems* 46(1): 254–264.

Kumaraguru P, Sheng S, Aquisti A, Cranor LF, Hong J (2010) Teaching Johnny not to fall for phish. *ACM Trans. Internet Tech.* 10(2):1–31.

Lennon M (2011) Cisco: Targeted attacks cost organizations $1.29 billion annually. *Security Week* (June 30). Accessed August 14, 2016, https://www.securityweek.com/cisco-targeted-attacks-cost-organizations-129-billion-annually.

Li L, Helenius M (2007) Usability evaluation of anti-phishing toolbars. *J. Comput. Virology* 3(2):163–184.

Li L, Berki E, Helenius M, Ovaska S (2014) Toward a contingency approach with whitelist-and blacklist-based anti-phishing applications: What do usability tests indicate? *Behav. Inform. Tech.* 33(11):1136–1147.

Li S, Schmitz R (2009) *A Novel Anti-Phishing Framework Based on Honeypots* (IEEE, New York).

Liang H, Xue Y (2009) Avoidance of information technology threats: A theoretical perspective. *Management Inform. Systems Quart.* 33(1):71–90.

Liu W, Deng X, Huang G, Fu AY (2006) An antiphishing strategy based on visual similarity assessment. *IEEE Internet Comput.* 10(2):58–65.

Ma Z, Sheng ORL, Pant G, Iriberri A (2012) Can visible cues in search results indicate vendors' reliability? *Decision Support Systems* 52(3):768–775.

Mahmood MA, Siponen M, Straub D, Rao HR, Raghu TS (2010) Moving toward black hat research in information systems security: An editorial introduction to the special issue. *Management Inform. Systems Quart.* 34(3):431–433.

Mayer RC, Davis JH, Schoorman FD (1995) An integrative model of organizational trust. *Acad. Management Rev.* 20(3):709–734.

McAfee (2013) McAfee threats report. *First quarter*. McAfee (April 10). Accessed March 23, 2017, http://www.mcafee.com/us/resources/reports/rp-quarterly-threat-q1-2013.pdf.

McCullagh P (1980) Regression models for ordinal data. *J. Royal Statist. Soc. B* 42(2):109–127.

McKnight DH, Choudhury V, Kacmar C (2002) Developing and validating trust measures for e-commerce: An integrative typology. *Inform. Systems Res.* 13(3):334–359.

McKnight DH, Cummings LL, Chervany NL (1998) Initial trust formation in new organizational relationships. *Acad. Management Rev.* 23(3):473–490.

Morris M, Venkatesh V, Ackerman P (2005) Gender and age differences in employee decisions about new technology: An extension to the theory of planned behavior. *IEEE Trans. Engrg. Management* 52(1):69–84.

Musthaler L (2013) Security analytics will be the next big thing in IT security. *Network World* (May 31). Accessed March 23, 2017, https://www.networkworld.com/article/2166806/security-analytics-will-be-the-next-big-thing-in-it-security.html.

Oliveira D, Rocha H, Yang H, Ellis D, Dommaraju S, Weir D, Muradoglu M, et al. (2017) Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. *Proc. 2017 CHI Conf. on Human Factors in Computing Systems* (ACM, New York), 6412–6424.

Parrish JL Jr, Bailey JL, Courtney JF (2009) *A Personality Based Model for Determining Susceptibility to Phishing Attacks* (University of Arkansas, Little Rock).

Pavlou PA, Gefen D (2004) Building effective online marketplaces with institution-based trust. *Inform. Systems Res.* 15(1):37–59.

Porter CE, Donthu N (2006) Using technology acceptance model to explain how attitudes determine Internet usage: The role of perceived access barriers and demographics. *J. Bus. Res.* 59:999–1007.

Prat N, Comyn-Wattiau I, Akoka J (2015) A taxonomy of evaluation methods for information systems artifacts. *J. Management Inform. Systems* 32(3):229–267.

Qi X, Davison BD (2009) Web page classification: Features and algorithms. *ACM Comput. Survey* 41(2):1–31.

Rajab M, Ballard L, Jagpal N, Mavrommatis P, Nojiri D, Provos N, Schmidt L (2011) Trends in circumventing web-malware detection. Technical report, Google, https://static.googleusercontent.com/media/research.google.com/en//archive/papers/rajab-2011a.pdf, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.357.2542.

Ransbotham S, Mitra S (2009) Choice and chance: A conceptual model of paths to information security compromise. *Inform. Systems Res.* 20(1):121–139.

Richards K, LaSalle R, van den Dool F, Kennedy-White J (2017) 2017 cost of cyber crime study. Technical report, Ponemon Institute, North Traverse City, MI, https://www.accenture.com/_acnmedia/PDF-62/Accenture-2017CostCybercrime-US-FINAL.pdf#zoom=50.

Rogers RW, Prentice-Dunn S (1997) Protection motivation theory. Gochman DS, ed. *Handbook of Health Behavior Research 1: Personal and Social Determinants* (Plenum Press), 113–132.

Santhanam R, Sethumadhavan M, Virendra M (2010) *Cyber Security, Cyber Crime and Cyber Forensics: Applications and Perspectives* (Information Science Reference/IGI Global, Hershey, PA).

Sarkar S, Vance A, Ramesh B, Demestihas M, Wu D (2020) The influence of professional subculture on information security policy violations: A field study in a healthcare context. *Inform. Systems. Res.* 31(4):1240–1259

Schneier B (2000) Inside risks: Semantic network attacks. *Commun. ACM* 43(12):168.

Shashua A, Levin A (2003) Ranking with large margin principle: Two approaches. *Adv. Neural Inform. Processing Systems* 15: 961–968.

Sheng S, Holbrook M, Kumaraguru P, Cranor LF, Downs J (2010) Who falls for phish?: A demographic analysis of phishing susceptibility and effectiveness of interventions. *Proc. SIGCHI Conf. on Human Factors in Computing Systems* (ACM, New York), 373–382.

Shields K (2015) Cybersecurity: Recognizing the risk and protecting against attacks. *NC Banking Inst.* 19:345.

Shmueli G, Koppius O (2011) Predictive analytics in information systems research. *Management Inform. Systems Quart.* 35(3): 553–572.

Siponen M, Vance A (2010) Neutralization: New insights into the problem of employee information systems security policy violations. *Management Inform. Systems Quart.* 34(3):487–502.

Sunshine J, Egelman S, Almuhimedi H, Atri N, Cranor LF (2009). Crying wolf: An empirical study of SSL warning effectiveness. *Proc. USENIX Security Sympos.* (USENIX Association, Berkeley, CA), 399–416.

Symantec (2012) Norton cybercrime report: The human impact. Symantec (April 10). Accessed March 23, 2017, http://us.norton.com/content/en/us/home_homeoffice/media/pdf/cybercrime_report/Norton_USA-Human%20Impact-A4_Aug4-2.pdf.

Taylor B (2014) How big data are changing the security analytics landscape. *TechRepublic* (January 2). Accessed March 23, 2017, https://www.techrepublic.com/blog/big-data-analytics/how-big-data-is-changing-the-security-analytics-landscape/.

Vance A, Lowry PB, Eggett D (2015) Increasing accountability through user-interface design artifacts: A new approach to address the problem of access-policy violations. *Management Inform. Systems Quart.* 39(2):345–366.

Vance A, Anderson B, Kirwan B, Eargle D (2014) Using measures of risk perception to predict information security behavior: Insights from electroencephalography (EEG). *J. Assoc. Inform. Systems* 15(10):679–722.

Vance A, Jenkins J, Anderson B, Bjornn D, Kirwan B (2018) Tuning out security warnings: A longitudinal examination of habituation through fMRI, eye tracking, and field experiments. *Management Inform. Systems Quart.* 42(2):355–380.

Venkatesh V, Morris M, Davis G, Davis F (2003) User acceptance of information technology: Toward a unified view. *Management Inform. Systems Quart.* 27(3):397–423.

Verizon (2016) Data breach investigations report. Accessed March 23, 2017, http://www.verizonenterprise.com/DBIR/2016/

Vishwanath A, Herath T, Chen R, Wang J, Rao HR (2011) Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems* 51(3):576–586.

Wang DY, Savage S, Voelker GM (2011) Cloak and dagger: Dynamics of web search cloaking. *Proc. 18th ACM Conf. on Computer and Communications Security* (ACM, New York), 477–490.

Wang J, Gupta M, Rao HR (2015) Insider threats in a financial institution: Analysis of attack-proneness of information systems applications. *Management Inform. Systems Quart.* 39(1):91–112.

Wang J, Li Y, Rao HR (2016) Overconfidence in phishing email detection. *J. Assoc. Inform. Systems* 17(11):759.

Wang J, Li Y, Rao HR (2017) Coping responses in phishing detection: An investigation of antecedents and consequences. *Inform. Systems Res.* 28(2):378–396.

Wang J, Chen R, Herath T, Vishwanath A, Rao HR (2012) Phishing susceptibility: An investigation into the processing of a targeted spear phishing email. *IEEE Trans. Professional Comm.* 55(4): 345–362.

Wright RT, Marett K (2010) The influence of experiential and dispositional factors in phishing: An empirical investigation of the deceived. *J. Management Inform. Systems* 27(1):273–303.

Wright RT, Jensen ML, Thatcher JB, Dinger M, Marett K (2014) Influence techniques in phishing attacks: An examination of vulnerability and resistance. *Inform. Systems Res.* 25(2):385–400.

Wu M, Miller RC, Garfunkel SL (2006) Do security toolbars actually prevent phishing attacks? *Proc. SIGCHI Conf. on Human Factors in Computing Systems* (ACM, New York), 601–610.

Zahedi FM, Song J (2008) Dynamics of trust revision: Using health infomediaries. *J. Management Inform. Systems* 24(4):225–248.

Zahedi FM, Abbasi A, Chen Y (2015) Fake-website detection tools: Identifying elements that promote individuals' use and enhance their performance. *J. Assoc. Inform. Systems* 16(6):448.

Zhang D, Yan Z, Jiang H, Kim T (2014) A domain-feature enhanced classification model for detection of phishing e-business websites. *Inform. Management* 51(7):845–853.

Zhang Y, Egelman S, Cranor L, Hong J (2007) Phinding phish: Evaluating anti-phishing tools. *Proc. 14th Annual Network and Distributed System Security Sympos*. 1–16.