Check for updates

Embedding Training Within Warnings Improves Skills of Identifying Phishing Webpages

Aiping Xiong, **Robert W. Proctor**, **Weining Yang**, and **Ninghui Li**, Purdue University, Lafayette, Indiana, USA

Objective: Evaluate the effectiveness of training embedded within security warnings to identify phishing webpages.

Background: More than 20 million malware and phishing warnings are shown to users of Google Safe Browsing every week. Substantial click-through rate is still evident, and a common issue reported is that users lack understanding of the warnings. Nevertheless, each warning provides an opportunity to train users about phishing and how to avoid phishing attacks.

Method: To test use of phishing-warning instances as opportunities to train users' phishing webpage detection skills, we conducted an online experiment contrasting the effectiveness of the current Chrome phishing warning with two training-embedded warning interfaces. The experiment consisted of three phases. In Phase 1, participants made login decisions on 10 webpages with the aid of warning. After a distracting task, participants made legitimacy judgments for 10 different login webpages without warnings in Phase 2. To test the long-term effect of the training, participants were invited back a week later to participate in Phase 3, which was conducted similarly as Phase 2.

Results: Participants differentiated legitimate and fraudulent webpages better than chance. Performance was similar for all interfaces in Phase 1 for which the warning aid was present. However, training-embedded interfaces provided better protection than the Chrome phishing warning on both subsequent phases.

Conclusion: Embedded training is a complementary strategy to compensate for lack of phishing webpage detection skill when phishing warning is absent.

Application: Potential applications include development of training-embedded warnings to enable security training at scale.

Keywords: cybersecurity, phishing, training, action on cybersecurity, procedural knowledge

Address correspondence to Aiping Xiong, College of Information Sciences and Technology, the Pennsylvania State University, E373 Westgate Building, University Park, PA 16802, USA; e-mail: axx29@ist.psu.edu

HUMAN FACTORS
Vol. 61, No. 4, June 2019, pp. 577–595
DOI: 10.1177/0018720818810942
Article reuse guidelines: sagepub.com/journals-permissions
Copyright © 2018, Human Factors and Ergonomics Society.

INTRODUCTION

Phishing is a social engineering attack that uses e-mail, social network webpages, and other media to communicate messages intended to persuade potential victims to perform certain actions or divulge confidential information for the attacker's benefit in the context of cybersecurity (Khonji, Iraqi, & Jones, 2013; Orgill, Romney, Bailey, & Orgill, 2004).

Because the website mimics that of a reputable organization, victims are tricked into entering personal information and credentials, which are then stolen by the attackers. Damages from phishing attacks include financial losses, exposure of privacy information, and reputational harm to companies. Phishing is estimated to have resulted in about $30 million in damages to U.S. consumers and businesses in 2017 (FBI, 2018). Beyond financial loss, users reported reduced trust in people and the technology as a consequence of phishing attacks (Kelley, Hong, Mayhorn, & Murphy-Hill, 2012).

Because of the negative consequences of phishing attacks, considerable effort has been devoted to devising methods to protect users from them. Detection and prevention of phishing scams is the first line of protection to stop attacks from reaching people. Computer scientists have developed several automated tools for phishing detection: (1) e-mail classification at server and client levels to filter phishing e-mails (e.g., Fette, Sadeh, & Tomasic, 2007); (2) website blacklists consisting of phishing URLs and IP addresses detected in the past (e.g., Google Safe Browsing; Whittaker, Ryner, & Nazif, 2010) or almost all possible variants of a URL (e.g., Prakash, Kumar, Kompella, & Gupta, 2010); (3) heuristic solutions based on sets of rules from previous real-time phishing attacks to detect zero-day (i.e., previously unknown) phishing attacks (e.g., Zhang, Hong, & Cranor, 2007); (4) webpage visual-similarity assessments to block phishing websites (e.g., Fu, Liu, & Deng, 2006). However, those tools and services do not protect against all phishing due to evolution of phishing attacks and the difficulty computers have in accurately extracting the meaning of the natural language messages in e-mails (Stone, 2007).

When automatic detection fails, the user makes the final decision on a webpage's legitimacy (Proctor & Chen, 2015). Thus, researchers developed decision-aid tools to warn users when a fraudulent website is detected. The tools include dynamic security skins (Dhamija & Tygar, 2005), browser toolbars (Herzberg & Gbara, 2004), and web browser phishing warnings and secure sockets layer (SSL) warnings (Carpenter, Zhu, & Kolimi, 2014; Felt et al., 2015). Those tools remind users of potential risks passively or actively. Passive warnings employ principles, such as colored icons or highlighting, which signal potential dangers to users without interrupting their primary tasks (Chou, Ledesma, Teraguchi, & Mitchell, 2004; Herzberg & Gbara, 2004; Lin, Greenberg, Trotter, Ma, & Aycock, 2011). Active warnings capture users' attention by forcing them to choose one of the options presented by the warnings (Egelman, Cranor, & Hong, 2008; Felt et al., 2015; Wu, Miller, & Garfinkel, 2006).

Yet, these decision-aid tools have evidenced ineffectiveness (e.g., Xiong, Proctor, Yang, & Li, 2017) and usability problems (e.g., Sheng et al., 2009; Wu et al., 2006). Specifically, people showed a lack of understanding of the decision-aid warnings in general (e.g., Felt et al., 2015; Wu et al., 2006). Training is one promising approach to address users' lack of comprehension, and a prior study provided evidence that knowledge gained from training enhanced the effectiveness of a phishing warning (Yang, Xiong, Chen, Proctor, & Li, 2017). Currently, there is little work on integrating phishing training and warning. We conjectured that such research is essential because of (a) the inability to require the large population of internet users to take classroom training, and (b) minimal warning protection for zero-day attacks.

Our aim in the current study was to understand the effect of embedded training within phishing warnings in helping users detect phishing webpages. We conducted an experiment to address three research questions:

1. What are the short- and long-term effects of training that is embedded within a phishing warning?
2. Which is the most effective way to present training to help users learn skills of how to identify the legitimacy of a webpage?
3. Does presenting training-embedded warnings as feedback of users' actions facilitate the effect of training?

ACTION-ORIENTED PHISHING PROTECTION STRATEGIES

Phishing Warning

When warnings were presented to aid users' decisions, users who clicked through the warnings showed a lack of understanding of the warnings (e.g., Bravo-Lillo, Cranor, Downs, & Komanduri, 2011; Dhamija, Tygar, & Hearst, 2006). These findings are somewhat unexpected because most of the warning designs followed guidelines to improve users' understanding of the risks, for example, using direct language and symbols to describe explicit consequences of the risk (Felt et al., 2015; Yang et al., 2017). Nevertheless, scrutiny of the information presented in those warnings revealed a focus on facts about phishing (e.g., the definition and potential costs), also known as *declarative* knowledge (Anderson, 2013).

Downs, Barbagallo, and Acquisti (2015) investigated differences between declarative knowledge about phishing and *procedural* knowledge of the actions to determine URL legitimacy (Anderson, 2013). In an online role-play study, participants chose possible actions for legitimate and fraudulent e-mails and possible actions for webpages following each e-mail's link. Declarative knowledge was closely related to participants' self-reported predictions on awareness, susceptibility and intentions, but procedural knowledge was the only predictor of the users' ability to adjust their risk decisions.

Xiong et al. (2017) conducted a study, in a laboratory setting with an eye-tracker, investigating why a passive warning (domain highlighting) is ineffective at helping users identify phishing webpages. They based their study on the fact that the domain name embedded within the URL of a phishing site will always be different from the legitimate one. Thus, the mismatch between the real domain name and the impersonated webpage serves as a reliable cue to detect phishing attacks (Lin et al., 2011). Because users may overlook the domain name (Jagatic, Johnson, Jakobsson, & Menczer, 2007), the domain of whichever site a user is currently viewing is highlighted. Specifically, the domain name portion within the URL in the browser's address bar is in black, whereas the rest of the URL is in gray (e.g., Google Chrome, Firefox).

In Xiong et al.'s (2017) study, participants evaluated the safety of legitimate and fraudulent webpages in two phases, with instructions to look at the address bar in the second phase but not initially. Although safety evaluation results showed some benefit of attending to the address bar, domain highlighting did not provide effective protection against phishing attacks. Yet eye-tracking results (e.g., heat map) revealed that participants' visual attention was attracted by the highlighted domains. Thus, the ineffectiveness of domain highlighting seems due to participants' lack of knowledge concerning how to use the domain name to identify webpage's legitimacy.

Equipping users with skills of how to identify potential phishing webpages (procedural knowledge) seems to be critical to improve the effectiveness of phishing warnings. Yang et al. (2017) investigated the effectiveness of phishing training and its interaction with a phishing warning on the webpage. The training content focused on how to evaluate the webpage's legitimacy by using the domain name. In a field experiment, participants in four groups varying in the presence and absence of the phishing training and warning received a simulated phishing e-mail attack targeting Amazon. Although, many participants who received only the training or only the warning fell prey to the simulated phishing attack, none of the participants who received both interventions submitted their genuine account information.

That no advantage was evident for the condition with only training indicates the necessity of making users aware of security issues through warnings when the issues arise, consistent with the idea that security typically is a secondary goal. The results obtained in the warning-only condition are similar to previous findings (e.g., Felt et al., 2015), suggesting that security awareness alone is not sufficient to protect users from phishing attacks. The power of using a combination of training and phishing warning to reduce the likelihood of being phished provided evidence that participants should not only be aware of the risks but also equipped with skills to take actions on the risks. Thus, it is critical to figure out a way to integrate phishing training and warning such that a large population of internet users can be trained effectively.

Phishing Training

Because phishing threats cannot be eliminated entirely through automated tools or users' compliance with phishing warnings, users necessarily must be trained about phishing attacks and how to avoid being phished. Despite training being an essential aspect of cybersecurity, it is the least popular approach (Hong, 2012).

The most basic approach to training is to post information about phishing online, as done by academic organizations, government organizations, nonprofit organizations, and companies. For example, Anti-Phishing Work Group (APWG) provides the STOP-THINK-CONNECT global cybersecurity education and awareness campaign to improve the public understanding of phishing. Although such education and advice can improve users' ability to avoid phishing attacks, most members of the public will not read them.

In a classroom setting, Anandpara, Dingman, Jakobsson, Liu, and Roinestad (2007) examined the effectiveness of a phishing training (i.e., FTC Consumer Alert) at a test with the portion of phishing trials varied from 25% to 100%. Forty participants identified legitimate and phishing e-mails before and after the training. Across two test phases, there was no correlation between the actual phishing e-mails and the number of phishing e-mails that participants identified. Thus, Anandpara et al. claimed that the traditional forms of education increase the level of fear or concern among users but not the ability to identify phishing scams.

Ferguson (2005) evaluated a contextual training approach, sending fake phishing e-mails to participants to explore their vulnerability to phishing attacks in the real world. The study tested participants' ability to detect phishing attacks in the first phase. In the second phase, participants received phishing training and a lecture in a classroom and were then tested. Participants' ability to identify phishing e-mails improved after the training (also see Dodge, Carver, & Ferguson, 2007).

Based on contextual training, Kumaraguru et al. (2007, 2009) designed and evaluated an e-mail embedded-training system called PhishGuru to avoid phishing attacks. Participants received simulated phishing e-mails, and a training page appeared whenever participants clicked on a phishing link in the e-mail. Users' immediate and long-term ability to identify phishing attacks improved after receiving embedded training of phishing e-mails in both laboratory and real-world settings. Most forms of security training take place in a classroom and give people few opportunities to test what they have learned. In contrast, embedded training teaches people within the specific context of use in which they would normally be attacked (Caputo, Pfleeger, Freeman, & Johnson, 2014; Kumaraguru et al., 2007, 2009; Kumaraguru, Sheng, Acquisti, Cranor, & Hong, 2010). Thus, among the alternative training methods, embedded training, designed to teach users critical information during their typical online interactions, is the most promising (Al-Daeef, Basir, & Saudi, 2017).

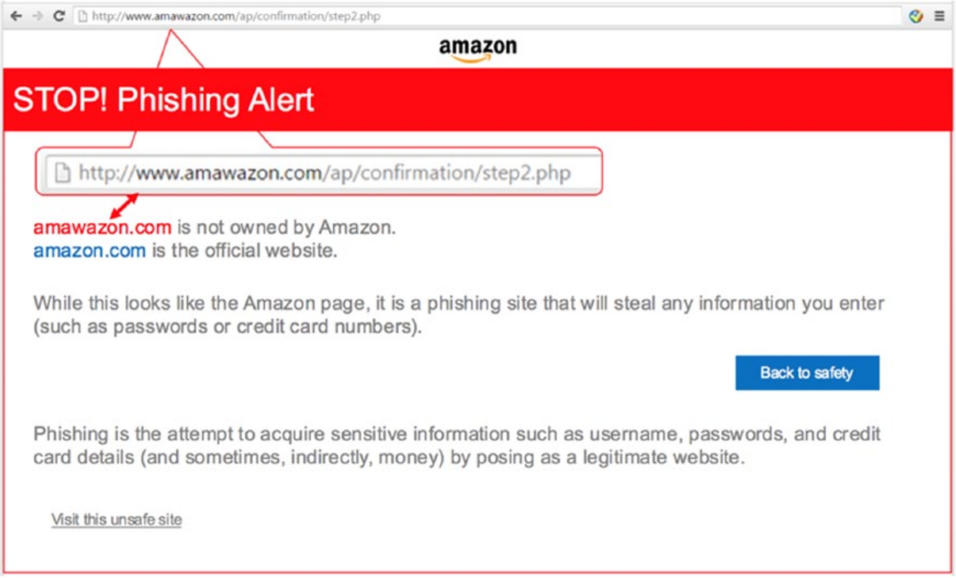
Nevertheless, previous work revealed that the potential effectiveness of embedded training is limited by the requirement that users read the training material. Kumaraguru et al. (2007, 2009) found that the training-embedded material is only effective when users actually read it, which they tend not to do if the training message is long (Caputo et al., 2014). Inspection of the training-embedded material used in prior studies, in fact, reveals long descriptions that require much time and effort. However, security is a user's secondary goal in general. Thus, it is critical to implement the training information in such a way that users can acquire and use it easily and quickly. Because warnings are present when users encounter potential phishing webpages, embedding training within phishing warning may be a good opportunity to equip users with skills for using the knowledge to regulate security-related behaviors.

Retention and Transfer of Knowledge Acquisition From Training

To detect zero-day phishing that has not been blacklisted or was missed by heuristic techniques, users need to retain the knowledge gained from training and transfer it to other situations. Retention is the ability of people to retrieve the concepts or procedures learned after a period of time. Transfer is the ability to apply the knowledge gained from one situation to another that differs from that of the knowledge acquisition (Roediger, Dudai, & Fitzpatrick, 2007). Both abilities are essential to phishing detection due to the thousands of new phishing URLs that are generated monthly (PhishTank, 2018).

The retention and transfer of knowledge is closely related to the process of acquisition, which is largely determined by the nature of the knowledge and how the knowledge is presented during training. Thus, first of all, one must be aware of which type of knowledge is involved during the training, namely, facts and events (declarative knowledge) or knowing how to do something (procedural knowledge). Also, to ensure effective and efficient training, a specific type of knowledge should be presented in line with its form of function or nature of representation. That is, declarative knowledge should be presented in a way that is available for recall or recognition, and procedural knowledge should be presented in a way that guides operations or actions by specifying what is to be done under which conditions (Oberauer, 2010).

Transfer and retention of declarative and procedural knowledge are widely accepted as having different properties (Healy & Bourne, 2012; Lee & Vakoch, 1996). Declarative knowledge declines quickly, whereas procedural knowledge, once acquired, remains at the same level when retested after one week or longer. Transfer is typically better for declarative knowledge. Yet, the continuous practice of procedures is accompanied by accumulative learning of factual information. Thus, retention and transfer of both types of knowledge is expected from training focused on procedural knowledge. Due to



http://www.amawazon.com/ap/confirmation/step2.php

amazon

STOP! Phishing Alert

http://www.amawazon.com/ap/confirmation/step2.php

amawazon.com is not owned by Amazon.
amazon.com is the official website.

While this looks like the Amazon page, it is a phishing site that will steal any information you enter (such as passwords or credit card numbers).

Back to safety

Phishing is the attempt to acquire sensitive information such as username, passwords, and credit card details (and sometimes, indirectly, money) by posing as a legitimate website.

Visit this unsafe site

Figure 1. Declarative training-embedded warning interface.

the procedural nature of using the domain name to identify phishing webpages, presenting the training content through step-by-step procedural instructions was expected to result in better acquisition and subsequent retention and/or transfer than presenting it with declarative sentences, i.e., descriptions of the legitimate and fraudulent domain names.

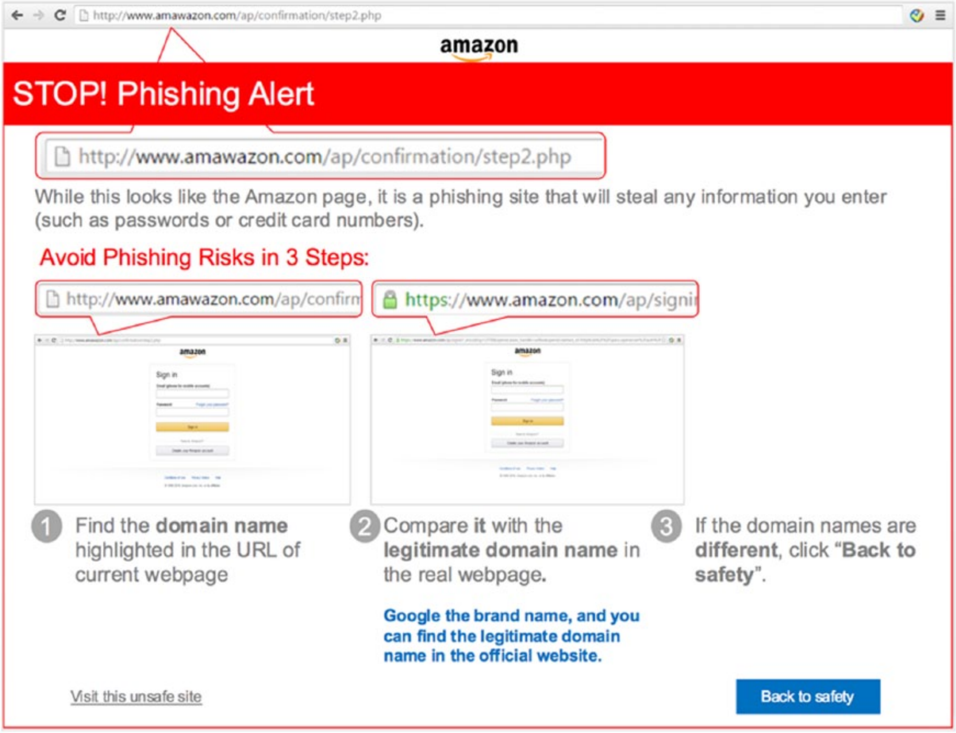
Action Effect

In most popular browsers, e.g., Google Chrome, after a user clicks the link within a phishing e-mail, the phishing warning blocks the whole webpage and any potential interactions with the webpage (Felt et al., 2015). But the anticipated consequence of an action can have an effect on the information processing that is required to initiate the action subsequently (Hommel, Müsseler, Aschersleben, & Prinz, 2001). Thus, the current warning implementation method may eliminate the possibility of users acquiring the knowledge and skills for phishing webpage detection. Instead of using the warning as a block to action, we proposed to implement the phishing warning as an immediate action effect, or feedback, to provide guidance toward correct behavior (Schmidt & Bjork, 1992).

PROPOSED TRAINING-EMBEDDED WARNING INTERFACES

We developed two new training-embedded warning interfaces, one we call Declarative (Figure 1) and the other Procedural (Figure 2). For both interfaces, a training intervention focusing on domain names is displayed within an active security warning to help users develop the knowledge and skills to detect potential phishing webpages. Due to focusing on the webpage's domain name, for both interfaces, a screenshot of the URL part is enlarged and linked to the URL by an arrow, indicating that this warning is specific for the domain name.

For the Declarative interface, the highlighted domain is listed below the URL screenshot, marked in red, and described as not owned by the related brand name with a sentence, such as "amawazon.com is not owned by Amazon." Because people have difficulty discriminating the credibility of websites based on domain names (Wogalter & Mayhorn, 2008), the legitimate domain name is also listed for comparison. The pairwise phishing and legitimate domain names serve as instances to train users about the domain-name spoof methods (the *similar* and *complex* methods used in current study, which we explain in the Method section).


http://www.amawazon.com/ap/confirmation/step2.php
amazon
http://www.amawazon.com/ap/confirm
https://www.amazon.com/ap/signi
Visit this unsafe site
Back to safety

STOP! Phishing Alert

http://www.amawazon.com/ap/confirmation/step2.php

While this looks like the Amazon page, it is a phishing site that will steal any information you enter (such as passwords or credit card numbers).

Avoid Phishing Risks in 3 Steps:

1. Find the **domain name** highlighted in the URL of current webpage
2. Compare **it** with the **legitimate domain name** in **the real webpage.**

   Google the brand name, and you can find the legitimate domain name in the official website.
3. If the domain names are **different**, click **"Back to safety"**.

Figure 2. Procedural training-embedded warning interface.

The embedded training content within the Procedural interface is the same as the Declarative one except that the two sentences are replaced by actions of how to avoid phishing risks in three steps: (1) Find the domain name highlighted in the URL; (2) Compare it with the legitimate domain name; (3) If the two domain names are different, click "Back to Safety." Note that the first two steps embed not only the pairwise domain names comparison as in the Declarative interface, but also how to get the information explicitly.

In compliance with the guidelines of efficient warning that Laughery and Wogalter (2006) proposed, for both interfaces the signal words "STOP! Phishing Alert" are included at the top to indicate the level of risk present. Both interfaces describe the nature of the phishing risk and its explicit consequences in a specific, complete, but not too lengthy text, to enhance the awareness of phishing. We used simple language to make the training message accessible to as many users as possible. In addition, the interface includes directive action (i.e., opinionated action; Felt et al., 2015), which is the "Back to Safety" button highlighted in blue. Participants had the opportunity to go back to the phishing webpage by the "Visit this unsafe site" button.

THE EXPERIMENT

Using a between-subject design, we studied six conditions involving two factors: warning interface and time at which the warning was presented. The three interfaces were: (a) Chrome: current Chrome phishing warning; (b) Declarative: the declarative training-embedded warning interface; (c) Procedural: the procedural training-embedded warning interface. Training alone was not included because we previously found that a training-only condition was ineffective (Yang et al., 2017). The warning was presented before the webpage (Before condition), which is the same as the current phishing warning presentation in Chrome, or after the webpage (After condition), in which case the warning was presented only when participants entered any information on phishing webpages. For the After condition, based on participants' responses to the two phishing webpages, the frequency with which they would see the warning was zero, once, or twice. Only participants who saw the warning at least once were included to examine the effect of warning interface. Because the resulting warning frequencies could reflect participants' prior and acquired knowledge, we also examined the three frequencies' results of the After condition and their interaction with the warning interfaces.

We conducted a three-phase experiment to evaluate the short- and long-term effectiveness of the two training-embedded warning interfaces against the control condition (Chrome). In Phase 1, participants made login decisions on legitimate and fraudulent webpages and got warning/training on phishing webpages. After a distraction task, participants judged webpages' legitimacy without warnings to evaluate the short-term effect of the embedded training in Phase 2. One week later, we invited each participant to return for Phase 3 to evaluate the legitimacy of extra webpages, again without warnings, to examine the long-term effect of the embedded training.

We predicted that the two training-embedded warning interfaces would yield better phishing-detection performance than the control condition, particularly in Phases 2 and 3 when there was no warning and participants could use the knowledge learned from the training to identify phishing webpages. We expected this effect of training to be more evident for the Procedural interface than for the Declarative interface because of the former's stepwise depiction of using domain names to identify the webpage's legitimacy. Finally, presenting the warning as an action effect rather than as a block to the webpage may be more effective for learning, which would yield better performance when warnings were absent.

Method

Participants. We recruited 1,080 participants (63% female) through Amazon Mechanical Turk (MTurk) in July and August 2016. In the Before condition, 120 participants received each of the three interfaces. For the After condition, because some participants would not see the warning based on the actions they selected, we doubled the number of participants to 240 for each interface. Approximately 120 participants saw the warning at least once for each interface. Participants' ages ranged from 18 to over 50 years, with 74% between 18 and 40 years. 91% were college students or professionals who had a bachelor's or higher degree. 83% of the participants claimed that they do not have a degree or work experience in computer science or similar fields. The demographic distributions between conditions were similar. Each participant was compensated $0.50 for Phases 1 and 2. A total of 639 participants (432 of whom saw warnings initially) returned for Phase 3. Return rates and demographics were similar across conditions, and participants who finished Phase 3 received an extra $0.25.

This experiment complied with the American Psychological Association Code of Ethics and was approved by the institutional review board at Purdue University. Informed consent was obtained from each participant. The experiment data that were stored and analyzed are anonymized.

Apparatus and stimuli. The study was performed with participants' own laptop or computer. To ensure the training content's readability, we did not allow participants to continue the study if they were using any mobile device. We limited data collection to participants from the United States because the websites used in the study are popular in this country.

The details of phishing and legitimate webpages for each phase are listed in Tables 1 and 2, respectively. Each webpage was an exact replica of the original website except the URL of each phishing webpage, which was a valid phishing URL listed in PhishTank. We included SSL for legitimate webpages as in the real world. This difference between legitimate and fraudulent sites was constant across conditions and phases and should have no differential impact on the comparison among the three warning interfaces. The six most-targeted phishing industries (see Table 1), such as bank and e-commerce, were selected. For each phase, phishing trials came from only two categories and were selected from the most popular websites within each category.

To evaluate retention of the embedded-training, the same two spoof methods were used across phases, which also made the difficulty of identifying phishing webpages equal. One is the *similar* method, in which fraudulent URLs are

TABLE 1: URLs (Spoofed, Original) of Phishing Webpages for Each Phase, Category, and Website

Phase	Category	Website	Spoofed URL	Original URL
Phase 1	E-commerce	Amazon	http://www.amawazon.com	https://www.amazon.com/ap/signin?_encoding=UTF8&openid.assoc_handle=usflex&…
		eBay	http://umpapa.lt/account999865…	https://signin.ebay.com/ws/eBayISAPI.dll?SignIn&ru=http% 3A%2F%2Fwww.ebay.com%2F
	Bank	Bank of America	http://www.arfcorretora.com.br/BofA/signon.php…	https://www.bankofamerica.com/sitemap/hub/signin.go
		Chase	http://www.tulsicomputers.com/system/logs/OnlineChase/	https://chaseonline.chase.com/
		Wells Fargo	http://plaskit.fr/ibraries/wellsfargo/wellsfargo/…	https://www.wellsfargo.com/
Phase 2	Social media	Facebook	http://info-setings.usite.pro/facebook-support.html…	https://www.facebook.com/
		Twitter	http://twiller.org	https://twitter.com/login?lang=en
	E-mail	Gmail	http://www.achyro89.com/google/business/google/Ed…	https://accounts.google.com/ServiceLogin?service=mail&passive=true&rm=false&continue…
		Microsoft	http://365-outlook.com-useronlineereset72.microsoftexchange1…	https://outlook.office.com/owa/#authRedirect=true
		Yahoo	http://www.assomabauru.org.br/Yahoo/Yahoo-2014/yinput...	https://login.yahoo.com/?.src=ym&.intl=us&.lang=en-US&.done=https%3a//mail.yahoo.com
Phase 3	Cloud storage	Apple	http://www.steaksmore.com/files/apple...	https://appleid.apple.com/#!&page=signin
	Government	IRS	http://irs.gov.irs-qus.com	https://www.irs.gov/refunds

TABLE 2: URLs of Legitimate Webpages for Each Phase and Website

Experiment	Website	URL
Phase 1	BestBuy	https://www-ssl.bestbuy.com/identity/signin?token=tid%3A792f2c17-7d57-11e6-a4b4-005056920f07
	Economist	http://www.economist.com/
	Expedia	https://www.expedia.com/user/signin?ckoflag=0
	Glassdoor	https://www.glassdoor.com/profile/login_input.htm
	Pinterest	https://www.pinterest.com/login/
	Alamo	https://www.alamo.com/en_US/car-rental/reservation/startReservation.html
	Dropbox	https://www.dropbox.com/login
	Walmart	https://www.walmart.com/account/login?tid=0&returnUrl=%2F
Phase 2	TripAdvisor	https://rentals.tripadvisor.com/login
	LinkedIn	https://www.linkedin.com/uas/login
	Skype	https://login.skype.com/login?message=signin_continue
	Budget	http://www.budget.com/budgetWeb/home/home.ex
	Southwest	https://www.southwest.com/flight/login
	Macy's	https://m.macys.com/account/signin
Phase 3	Ibis	https://www.ibis.com/gb/northamerica/index.shtiml
	Uber	https://login.uber.com/login
	Comcast	https://login.comcast.net/login?r=comcast.net&%=oauth...
	Fitbit	https://www.fitbit.com/login
	Priceline	https://www.priceline.com/dashboard/#/login
	Hilton	https://secure3.hilton.com/en/hh/customer/login/index.htm

visually similar to the legitimate URLs. The other is the *complex* method, in which fraudulent URLs expand the length of the legitimate ones to make interpretation of the URL difficult (Lin et al., 2011). Also, we used different webpages across phases to test the transfer effect of the training.

Three sets of 10 different webpages were used as stimuli. Phase 1 included eight legitimate trials from Table 2 and two phishing trials from the bank and e-commerce categories in Table 1, respectively. Phishing trials of Phase 2 were selected from social media and e-mail categories. The eight legitimate trials included the six listed in Table 2 and legitimate versions of the two phishing pages of Phase 1. For Phase 3, URLs for the two phishing trials are listed in Table 1, and the legitimate trials included the six listed in Table 2 and legitimate versions of the two phishing trials in Phase 2. In each phase, all trials were presented randomly, and the possible combinations of phishing trials were presented in approximately equal number.

Procedure. Participants were allowed to participate in only one of the six conditions. Each study started with a questionnaire about participants' daily online browsing experience, such as browsing time every day, online time distribution of different activities, etc. The questionnaire did not mention phishing, or any other cybersecurity concern.

After the questionnaire, Phase 1 started, which was designed based on Dhamija et al.'s (2006) study of users' ability to identify phishing websites. Participants were told to imagine that they had an account with one website (e.g., Chase), and they just received an e-mail from the website asking them to click on one link within the e-mail. Then, supposing they clicked on the link and were directed to a webpage, participants were asked to choose their immediate action on the webpage. Participants received 10 different login webpages (8 legitimate, 2 fraudulent), making binary decisions for each webpage (i.e., *Enter e-mail address and password* or

Leave or close the webpage). For each decision, we also asked participants how confident they were in their decision on a scale of 1 to 5 (1 = not confident at all; 5 = very confident). Warning was presented to help participants make an informed decision on phishing trials. We measured the viewing time of webpage/warning presentation and the corresponding decision.

After completing Phase 1, participants performed 24 trials of a Stroop color-identification task (MacLeod, 1991) as a cognitively demanding distraction, in which they responded with a left or right keypress to the color (red or green) of a congruent or incongruent color word (red or green). The distraction task took about 3 min. Then, in Phase 2, participants made legitimacy judgments (i.e., Legitimate or Phishing) for 10 different login webpages without warning. We changed the task from webpage login decisions to legitimacy judgments for two reasons: (a) Our primary interest was to evaluate whether participants had learned to discriminate phishing webpages from the embedded training; (b) Participants should be aware that this study was about phishing after Phase 1, and previous studies showed that informed participants were significantly better at discriminating between phishing and genuine e-mails than uninformed participants (Parsons, McCormac, Pattinson, Butavicius, & Jerram, 2015). We also measured participants' confidence rating for each decision and viewing time of each webpage.

After the judgment task, participants completed a questionnaire that asked for demographic information (e.g., age, gender, education, computer science related work experience). Additionally, the questionnaire asked participants to select a potential outcome of phishing from a list of four options, to check their comprehension of the warning. Participants also estimated their possibility of falling for a phishing attack before and after the study on a 5-point scale (1 definitely will not be phished; 5 will fall for phishing attack for sure).

Phase 3 was conducted a week after Phases 1 and 2. Each participant received an e-mail message inviting him/her to evaluate another 10 webpages' legitimacy as in Phase 2. After completing their legitimacy decisions for those webpages, participants were tested by choosing the legitimate URL from among another five spoofed phishing URLs.

Results

Over 68% of participants reported that they spent more than 2 hr online every day. They indicated spending 22% of the time on social media, 20% on work or study, 15% on e-mail, 14% on a search engine, and 10% on online shopping. These results and others from the initial questionnaire were similar across conditions.

We measured the selected decision, confidence rating, and webpage/warning viewing time of each participant for each webpage. In Phase 1, decisions were coded as accurate when participants responded "Enter e-mail address and password" on legitimate trials and "Back to safety" on warning interfaces for phishing trials. Choices of "Leave or close the webpage" on legitimate webpages and "Visit this unsafe site" on warning interfaces for phishing webpages were coded as inaccurate. For the After condition in Phase 1, warnings were presented when participants selected "Enter e-mail address and password" for phishing trials, and decisions were measured based on their final decisions. That is, if a participant chose to enter the ID and password on a phishing webpage but corrected the decision later on the warning, we counted it as a correct decision. For legitimacy decisions in Phases 2 and 3, choices were coded as accurate when participants selected "Legitimate" for legitimate trials and "Phishing" for phishing trials.

For each phase, the number of correct decisions for phishing trials and legitimate trials was determined for each participant and grouped as a function of warning presentation (Before, After) × warning interface (Chrome, Declarative, Procedural). We used signal detection theory methods that allow assessment of sensitivity (d') to phishing and response bias (c) (e.g., Canfield, Fischhoff, & Davis, 2016; Xiong et al., 2017) based on correct responses to phishing trials (hits) and incorrect responses to legitimate trials (false alarms). To accommodate hit rates and false-alarm rates of 0 or 1, a log-linear correction added 0.5 to the number of hits and 0.5 to the number of false alarms and 1 to the number of signals (phishing webpages) or noise (legitimate webpages; Canfield et al., 2016; Hautus, 1995). The d' values of log-linear corrected data underestimate the true d' values (Hautus, 1995), but differences across the warning conditions should reflect differences apparent in the raw accuracy data (see Table 3). The d' and c measures were submitted to analysis of variance (ANOVA) with warning presentation × warning interface, as was viewing times. Participants' confidence ratings were generally high and did not vary much across conditions, so we do not report the statistical test results in the text but list mean values of each condition in Table 4.

Phase 1: Effect of warning interface. Table 3 includes correct decision rates of phishing and legitimate trials of each condition collapsed cross participants, as well as means of signal-detection parameters for each condition. Table 4 provides the means of webpage viewing time and confidence rating.

Signal-detection parameters. The three interfaces showed similar sensitivity ($d'_{\text{chrome}} = 1.22$, $d'_{\text{declarative}} = 1.24$, $d'_{\text{procedural}} = 1.24$) and bias toward judging webpages as fraudulent ($c_{\text{chrome}} = -0.28$, $c_{\text{declarative}} = -0.28$, $c_{\text{procedural}} = -0.24$), Fs < 1.02. Whether the warning was presented Before ($d' = 1.27$, $c = -0.25$) or After ($d' = 1.20$, $c = -0.28$) the webpages' presentation showed no influence on participants' sensitivity or bias, Fs < 1.21. When making login decisions with the aid of warnings, participants demonstrated moderate detection ability, along with a bias toward judging webpages as fraudulent. The signal-detection parameters revealed that the two training-embedded interfaces were comparable to the Chrome warning.

Viewing times. The viewing-time measures for phishing trials differed across the three interfaces (see Table 4), $F_{(2,708)} = 34.61, p < .001, \eta_p^2 = .089$. Post-hoc Bonferroni analysis showed that all pairwise tests were significant (*ps* < .001). Viewing time was longest with the Procedural interface (14.8 s), intermediate with the Declarative interface (12.0 s), and shortest with the Chrome interface (9.5 s). The longer viewing times for the new interfaces imply that participants processed the extra embedded-training messages. Viewing time was longer for the After condition (13.9 s) than the Before condition (10.3 s), $F_{(1,708)} = 48.37$, $p < .001$, $\eta_p^2 = .064$, but this difference did not vary across the interfaces, $F < 1.0$.

Warning frequencies in the After condition. In the After condition, participants saw a warning only when they chose to enter information on a phishing webpage. Thus, for each interface, some participants never saw the warning, some saw the warning once (on either the first or second phishing trial), and some saw it twice. For participants who did not see the warning, their decision accuracy was 100% for phishing trials. We conducted ANOVAs on d' and c across the three frequencies (zero, once, twice) and the three warning interfaces. We did the same analysis for average viewing times.

Participants' sensitivities were similar across the three warning frequencies ($d'_{\text{zero}} = 1.18$, $d'_{\text{once}} = 1.14$, $d'_{\text{twice}} = 1.30$), $F_{(2,711)} = 1.64, p = .194, \eta_p^2 = 005$, and did not differ across the interfaces, $F_{(4,711)} = 1.18$, $p = .315$, $\eta_p^2 = .007$. Participants who did not see the warning showed similar sensitivity as those who saw the warning, indicating their awareness and knowledge of phishing scams without any aid. Bias toward judging webpages as fraudulent differed across frequencies ($c_{\text{zero}} = -0.38$, $c_{\text{once}} = -0.35$, $c_{\text{twice}} = -0.17$), $F_{(2,711)} = 13.57$, $p < .001$, $\eta_p^2 = .037$. Post-hoc comparisons showed that participants who saw the warning twice had less bias than those who saw it once and those who did not see the warning (ps < .001), which did not differ ($p = .715$). The bias was similar across interfaces and the difference across frequencies was similar among the three interfaces, *Fs* < 1.02.

Viewing times differed across the three warning frequencies, $F_{(2,\ 711)} = 48.29$, $p < .001$, $\eta_p^2 = .120$. Post-hoc pairwise comparisons were all significant, ps < .027, being longest for participants who saw the warning once (15.5 s), intermediate for those who saw the warnings twice (11.4 s), shortest for those who did not see the warning (9.5 s). There was an interaction of frequency × warning interface, $F_{(4,\ 711)} = 9.50$, $p < .001$, $\eta_p^2 = .051$. Participants who did not see the warning spent similar time across the three interfaces, but participants who saw the warnings spent longer time on the two training-embedded interfaces.

Participants who saw the warnings twice spent less time and showed less bias to judge the legitimate trials as phishing than participants who saw the warning once. This outcome suggests that participants who saw the warnings twice might not have processed the content of the warning as much as participants who saw the warning once.

For legitimate webpages, the viewing times differed across frequencies, $F_{(2,\ 711)} = 3.84$, $p = .022$, $\eta_p^2 = .011$. Pairwise comparisons showed that participants who saw the warning once spent

TABLE 3: Mean Decision Results for Each Condition

			Phase 1					Phase 2				Phase 3 (1-week later)				
Warning Presentation	Warning Frequency	Warning Interface	Subjects No.	Phishing Trials	Legitimate Trials	*d′*	*c*	Phishing Trials	Legitimate Trials	*d′*	*c*	Returned Subjects No.	Phishing Trials	Legitimate Trials	*d′*	*c*
Before	2	Chrome	120	97.1%	64.0%	1.28	–0.27	58.8%	91.0%	1.40	0.53	77	66.9%	83.9%	1.28	0.31
		Declarative	120	95.4%	63.2%	1.23	–0.26	77.9%	88.1%	1.65	0.29	80	76.9%	85.8%	1.54	0.25
		Procedural	120	94.6%	65.7%	1.29	–0.22	75.4%	88.9%	1.64	0.33	72	82.6%	85.4%	1.65	0.19
After	2	Chrome	52	92.3%	66.1%	1.28	–0.18	66.3%	88.9%	1.47	0.42	28	67.9%	83.0%	1.27	0.29
		Declarative	42	95.2%	70.5%	1.47	–0.14	56.0%	89.9%	1.31	0.54	24	68.8%	78.1%	1.15	0.21
		Procedural	38	89.5%	64.1%	1.15	–0.19	67.1%	88.8%	1.48	0.41	25	66.0%	84.0%	1.25	0.31
	1	Chrome	66	97.7%	55.3%	1.06	–0.40	67.4%	86.7%	1.41	0.37	35	67.1%	87.1%	1.37	0.36
		Declarative	78	99.4%	57.1%	1.15	–0.38	76.3%	86.4%	1.58	0.28	44	83.0%	83.2%	1.56	0.14
		Procedural	78	96.2%	61.9%	1.21	–0.29	82.7%	88.9%	1.80	0.27	47	80.9%	84.8%	1.59	0.20
	0	Chrome	122	100%	60.9%	1.23	–0.35	87.7%	87.3%	1.84	0.19	73	88.4%	83.6%	1.69	0.10
		Declarative	120	100%	58.1%	1.16	–0.39	91.7%	88.0%	1.94	0.16	68	87.5%	82.2%	1.65	0.10
		Procedural	124	100%	57.2%	1.14	–0.40	86.7%	86.9%	1.81	0.19	66	85.6%	83.5%	1.63	0.13

Note. Subjects number (No.), percentage of correct decisions of phishing and legitimate trials, signal-detection parameters (*d′* and *c*) by warning presentation (Before, After), warning frequency, and warning interface (Chrome, Declarative, Procedural) for each phase.

TABLE 4: Mean Viewing Time and Confidence Rating of Each Condition

				Phase 1			Phase 2		Phase 3 (1-week later)		
Trial Type	Warning Presentation	Warning Frequency	Warning Interface	Subjects No.	Viewing Time (s)	Confidence Rating	Viewing Time (s)	Confidence Rating	Returned Subjects No.	Viewing Time (s)	Confidence Rating
Phishing	Before	2	Chrome	120	7.9	4.7	8.3	4.3	77	14.5	4.2
			Declarative	120	10.1	4.5	8.8	4.4	80	12.5	4.1
			Procedural	120	12.9	4.6	8.8	4.5	72	10.6	4.3
	After	2	Chrome	52	10.0	4.4	8.4	4.2	28	12.2	4.0
			Declarative	42	10.7	4.5	10.7	4.1	24	11.9	4.2
			Procedural	38	14.2	4.3	8.6	4.2	25	14.4	4.3
		1	Chrome	66	12.0	4.6	11.6	4.3	35	14.4	4.2
			Declarative	78	15.7	4.5	9.8	4.3	44	14.2	4.2
			Procedural	78	18.2	4.5	9.7	4.4	47	11.1	4.1
		0	Chrome	122	10.7	4.7	8.0	4.5	73	11.2	4.4
			Declarative	120	9.3	4.8	7.8	4.4	68	11.7	4.5
			Procedural	124	8.3	4.7	7.7	4.5	66	11.6	4.4
Legitimate	Before	2	Chrome	120	9.3	4.3	8.0	4.3	77	14.7	4.1
			Declarative	120	8.9	4.3	8.7	4.3	80	13.1	4.1
			Procedural	120	10.4	4.4	10.0	4.4	72	13.1	4.2
	After	2	Chrome	52	8.0	4.3	8.6	4.2	28	11.3	4.0
			Declarative	42	11.1	4.3	7.7	4.3	24	10.7	4.1
			Procedural	38	7.9	4.3	8.6	4.2	25	15.0	4.1
		1	Chrome	66	8.9	4.4	9.9	4.3	35	12.3	4.3
			Declarative	78	9.1	4.3	9.6	4.2	44	17.8	4.0
			Procedural	78	8.6	4.4	8.4	4.3	47	11.8	4.2
		0	Chrome	122	10.7	4.2	10.0	4.3	73	14.6	4.1
			Declarative	120	9.8	4.4	9.7	4.3	68	15.2	4.2
			Procedural	124	11.0	4.4	9.6	4.2	66	13.6	4.0

Note. Subjects number (No.), mean viewing time, and confidence rating by trial type (Phishing, Legitimate), warning presentation (Before, After), warning frequency, and warning interface (Chrome, Declarative, Procedural) for each phase.

less time (8.8 s) than participants who did not see the warning (10.5 s), $p = .037$, suggesting participants who did not see the warning may develop the habit of checking a webpage's legitimacy.

Phase 2: Short-term effect of embedded training. Participants judged the legitimacy of another 10 webpages without warnings being presented for the two phishing webpages. See results in Tables 3 and 4.

Signal-detection parameters. Participants' sensitivity to phishing webpages differed across the three interfaces, $F_{(2,708)} = 3.87$, $p = .021$, $\eta_p^2 = .011$. Post-hoc comparisons indicated that the sensitivity for the Procedural condition ($d' = 1.67$) was larger than that of the Chrome condition ($d' = 1.42$), $p = .016$, but not significantly different from that of the Declarative condition ($d' = 1.57$), $p = .526$. The difference between the Chrome and the Declarative conditions was not significant, $p = .210$. Whether the warning was presented Before or After the webpages' presentation did not influence participants' sensitivity ($d'_{\text{before}} = 1.57$, $d'_{\text{after}} = 1.54$), $F < 1.0$.

Positive c values indicate that participants had a bias to identify webpages as safe when there was no warning present. Although participants showed similar bias regardless of whether the warning was presented Before or After phishing webpages ($c_{\text{before}} = 0.38$, $c_{\text{after}} = 0.36$), $F < 1.0$, the bias was smaller for the two training-embedded conditions ($c_{\text{declarative}} = 0.33$, $c_{\text{procedural}} = 0.32$) than for the Chrome condition ($c_{\text{chrome}} = 0.46$), $F_{(2,708)} = 7.43$, $p = .001$, $\eta_p^2 = 021$. Moreover, the benefit for the two training-embedded interfaces was evident in the Before condition ($c_{\text{chrome}} = 0.53$, $c_{\text{declarative}} = 0.29$, $c_{\text{procedural}} = 0.33$) but not the *After* condition ($c_{\text{chrome}} = 0.39$, $c_{\text{declarative}} = 0.37$, $c_{\text{procedural}} = 0.31$), $F_{(2,708)} = 3.85$, $p = .022$, $\eta_p^2 = .011$.

Larger sensitivity but less bias provides evidence for the short-term effect of the two training-embedded warnings. The benefit for the Procedural interface was suggested by the best sensitivity across the three interfaces.

Viewing times. Participants spent a similar amount of time on phishing webpages regardless of which warning interface was presented initially or when the warning was presented (see Table 4), Fs < 2.78. For legitimate webpages, participants spent a similar amount of time regardless of interfaces or when the warning was presented, Fs < 1.0. However, the 2-way interaction of interface × Before/After warning presentation was significant, $F_{(2,708)} = 4.14$, $p = .016$, $\eta_p^2 = .012$. Viewing times did not differ in the After condition across Chrome, Declarative, and Procedural conditions (9.3 s, 9.0 s, 8.5 s), $F < 1.0$, but increased in the Before condition (8.0 s, 8.7 s, 10.0 s), $F_{(2,357)} = 3.55$, $p = .030$, $\eta_p^2 = .010$.

Warning frequencies in the After condition. Participants' sensitivity differed across the three frequencies ($d'_{\text{zero}} = 1.86$, $d'_{\text{once}} = 1.61$, $d'_{\text{twice}} = 1.42$), $F_{(2,711)} = 12.19$, $p < .001$, $\eta_p^2 = .033$. Post-hoc analysis revealed that participants who did not see warnings showed greater sensitivity than those who saw the warning once ($p = .003$) and twice ($p < .001$), which did not differ, $p < .209$. The bias to judge webpages as legitimate differed across frequencies as well, $F_{(2,711)} = 23.19$, $p < .001$, $\eta_p^2 = .061$. There were differences between each pair, ps < .002, with bias being smallest for participants who never saw a warning ($c = 0.18$), intermediate for those who saw the warning once ($c = 0.30$), and largest for those who saw it twice ($c = 0.45$). Warning interface did not show a main effect nor interact with warning frequency for both measures, $Fs < 1.35$.

After the distraction task, the average viewing time for phishing webpages differed across the three frequencies, $F_{(2,711)} = 5.14$, $p = .006$, $\eta_p^2 = .014$. Post-hoc comparisons showed that participants who saw the warning once in Phase 1 spent longer viewing time (10.3 s) than participants who did not see it (7.8 s), $p = .005$, but neither differed significantly from that for participants who saw the warning twice (9.2 s; ps > .462). Viewing times for legitimate webpages approached significance across frequencies, $F_{(2,711)} = 2.81$, $p = .061$, $\eta_p^2 = .008$ (zero = 9.8 s; once = 9.3 s, twice = 8.3 s).

Post-session questionnaire. For their estimates of falling for a phishing attack, participants' ratings of 1 (definitely not falling for phishing) increased by 12.4% after the first two phases, mainly due to a change of the ratings of 2 and 3 to rating of 1. The increases of not-falling-for-phishing were similar for the Before and After conditions (12% vs. 13%), across the three interfaces (Chrome vs. Declarative vs. Procedural: 10% vs. 14% vs. 13%), and, in the After condition, between the two warning frequencies (once vs. twice: 14% vs. 12%). With regard to a potential outcome of a phishing attack, 84% of participants chose the correct answer (i.e., Someone may steal your credit card number and make bad charges). The results were similar regardless of when the warning was presented (Before vs. After: 84% vs. 84%), the interface seen (Chrome vs. Declarative vs. Procedural: 82% vs. 86% vs. 84%), and the warning frequencies for the After condition (once vs. twice: 81% vs. 86%).

Phase 3: Long-term effect of embedded training. The results of Phase 3 (see Tables 3 and 4) were analyzed similarly as Phase 2. Only 59% of people from Phase 1 participated. We compared the results of Phases 1 and 2 for participants who returned with those who did not, and the pattern of results was similar.

Signal-detection parameters. After one week, participants' detectability still differed across interfaces ($d'_{\text{chrome}} = 1.30$, $d'_{\text{declarative}} = 1.49$, $d'_{\text{procedural}} = 1.56$), $F_{(2,426)} = 3.03$, $p = .049$, $\eta_p^2 = .015$. Post-hoc analysis revealed that only the difference between the Chrome and the Procedural conditions was significant ($p = .048$), but not the other two pairs (ps > .215), suggesting a benefit of the Procedural interface. Participants' bias to identify webpages as safe also differed across interfaces, $F_{(2,426)} = 3.41$, $p = .034$, $\eta_p^2 = .016$. Bias of the Chrome condition ($c = 0.32$) was larger than that of the Declarative ($c = 0.21$, $p = .054$) and the Procedural ($c = 0.22$, $p = .068$) conditions, which did not differ ($p = .997$). Whether the warning was presented before or after the webpages showed no impact on participants' sensitivity ($d'_{\text{before}} = 1.49$, $d'_{\text{after}} = 1.41$) or bias ($c_{\text{before}} = 0.25$, $c_{\text{after}} = 0.24$), $Fs < 1.0$. It did not interact with interface either, Fs < 1.10.

Viewing times. One week after training, for both phishing and legitimate trials, viewing times were similar regardless of interfaces or when the warning was presented, Fs < 1.84.

Warning frequencies in the After condition. The main effect of frequency was also significant after one week, $F_{(2,401)} = 6.87$, $p = .001$, $\eta_p^2 = .033$. The sensitivity of participants who saw the warning once ($d' = 1.52$) was similar to those who did not see warnings in Phase 1 ($d' = 1.66$), $p = .337$, indicating an effect of embedded training. However, participants who saw the warning twice continued to show less sensitivity discriminating noise and signal ($d' = 1.23$) than participants in the other two conditions, ps < .055. After one week, participants' bias to judge webpages as safe differed across frequencies as well, $F_{(2,401)} = 7.37$, $p < .001$, $\eta_p^2 = .034$. Participants who did not see warnings still showed less bias ($c = 0.11$) than those who saw warnings once ($c = 0.22$, $p = .014$) and twice ($c = 0.27$, $p = .002$), which did not differ ($p = .638$). The effect of frequency did not interact with interface for both measures, Fs < 1.32.

Participants spent similar time on phishing webpages across the three frequencies $F_{(2,401)} = 1.93$, $p = .165$, $\eta_p^2 = .005$. For legitimate webpages, viewing time did not show any difference across frequencies either, Fs < 2.49.

Post-experiment questionnaire. About 43% of the participants chose the correct answer *www.pages.ebay.com/community/index.html* among the other five spoofed phishing URLs. The correct decision rates were similar for the Before (43%) and After (36%) conditions. Correct decision rate tended to be larger for the Procedural condition (46%) than for the Declarative (32%) or Chrome (37%) condition, $\chi^2_{(2)} = 5.29$, $p = .070$. For the After condition, correct decision rates were the same 36% for the two warning frequencies.

The top two incorrect answers were *www.ebay.com.ebay-billing.us/login* (43%) and *www.goecities.com/www.paypal.com* (10%), both of which used a spoof method that is different from those used in the current study. For the most-selected wrong answer, the selection rates were similar for the Before and After conditions (46% vs. 42%). Across the three interfaces, the Procedural condition showed the smallest error rate (36%), followed by Chrome (45%) and the Declarative (50%) conditions. For the After condition, error rates were 42% for both warning frequencies. Only 4% of participants chose one of the spoofed phishing URLs that were similar to the methods used in training (*www.account-verifyication.com/ebay/verify, www.147.46.236.66/paypal/login.html*, or *www.paypa1.com*).

DISCUSSION

We proposed two training-embedded warning interfaces and evaluated their effects in helping users to identify phishing webpages. The signal-detection analyses of Phase 1 showed that the two new interfaces were comparable to the current Chrome warning. When no warnings were displayed in Phases 2 and 3, larger sensitivity and smaller bias were obtained for the training-embedded interfaces, most clearly the Procedural interface. Together with the longer viewing time of the two warning-embedded interfaces at Phase 1, these results suggest that participants processed the training messages. Participants improved their ability to identify fraudulent webpages from viewing the training-embedded warning interfaces in both short-term and long-term (see also Kumaraguru et al., 2007, 2009), indicating that participants can retain and transfer the knowledge gained from the embedded training despite a limited opportunity for training. Furthermore, participants' improved performance was obtained without the expense of extra time to identify the phishing webpages, suggesting that it is efficient to use the domain name to identify phishing webpages.

Procedural vs. Declarative Interfaces

By using different website categories in evaluating the short- and long-term effects of warning interfaces, the retention and transfer of knowledge gained from both embedded-training interfaces was evident. Moreover, the Procedural interface showed better sensitivity at identifying phishing webpages in both short term and long term compared with the Chrome interface. However, different from our expectations, the Procedural interface showed only numerically better sensitivity than the Declarative interface, which may be due to there being only two critical steps in identifying a webpage's legitimacy based on the domain name. A benefit of the Procedural interface was also implied by the test of identifying phishing URLs one week later. Participants who saw the Procedural interface tended to select more correct answers and fewer incorrect answers, indicating better transfer of the declarative knowledge (domain names) by using the stepwise instructions.

Over 40% of participants mistakenly selected URLs as correct that employed a spoof method different from those used in the current study. This result suggests that the transfer benefit evident in Phases 2 and 3 may be restricted to the specific spoof methods, also termed near-transfer effects (Perkins & Salomon, 1992). Therefore, in practice, different spoof methods need to be implemented in training-embedded warnings to give users more varied training opportunities.

Action Effect

When warnings were presented as a consequence of users' action selection, the effect of the embedded-training was evident for participants who saw the warning once but not those who saw it twice. In the After condition, participants who were knowledgeable about phishing scams never saw the warning and showed better performance than participants who saw the warning once or twice in Phase 1. However, one week later, participants who saw the warning once showed a similar detectability of phishing webpages as the knowledgeable participants, without spending longer time. Also, their performance was better than that of the participants who saw the warning twice. Therefore, participants who saw the warning once acquired knowledge from the training message, whereas participants who saw the warning twice did not learn much.

Compared with the After condition, a benefit of the Before condition is that all users see the embedded training and get a training opportunity. Also, presenting the warning ahead has been shown to capture users' attention initially (Wogalter et al., 1987), which should increase the likelihood of users processing the training message. Presenting the warning before or after the webpage did not have a significant impact on the correct decision rate on phishing webpages in general. Thus, even when all participants were exposed to two training opportunities in the Before condition, some participants may have learned from the interfaces and some may have not. Future work is needed to investigate what factors contribute to the difference between participants.

Limitations

We used an experimental research method, manipulating warning interfaces and when the warning was presented, to obtain webpage legitimacy decisions in different phases. We are aware of studies using more ecologically valid methods (e.g., Felt et al., 2015), but we decided to present screenshots to exclude extraneous variables that may have an effect on the outcomes. By doing this, we are confident about the internal validity of the obtained results. Because the question of how far a study's results can be generalized to the real world is important, it is essential to evaluate training-embedded effectiveness in more naturalistic settings. Another possible confound was that the two new interfaces are novel, whereas participants may have experienced the Chrome interface previously. Although the novelty effect may play a role in Phase 1, better performance of the two new interfaces was evident without warnings afterwards. Finally, our participants were highly educated and young, so generalizing the findings to other user populations needs to be further examined.

Practical Implications

This study extended prior research about embedded training and showed a proof-of-concept for training-embedded warning. First, we showed the effectiveness of including training within a phishing warning, which provides a solution for implementing security training and reaching the general user population at scale. Second, instead of using lots of intensive training content, our embedded trainings are simple and short, focusing on the domain name, the most reliable cue for phishing detection. Such light-weighted training is easy for implementation and does not cost users much time and effort. Third, displaying training-embedded warnings before webpage presentation addresses the issue of users failing to attend to the training message. Fourth, both short- and long-term benefits of the Procedural interface suggest an advantage of the compatibility between training format and training content. Finally, our training-embedded warning interfaces validate the idea of combining different strategies to protect people from phishing attacks. The hybrid of warning/training suggests that security researchers and practitioners should consider combining different strategies to solve phishing and other cybersecurity issues.

ACKNOWLEDGMENTS

This research was supported by a National Security Agency Grant as part of a Science of Security Lablet through North Carolina State University and by the National Science Foundation under Grant No. 1314688.

KEY POINTS

- Embedding training within phishing warning addresses simultaneously two key factors impacting users' decisions of webpage legitimacy, awareness and action.
- A short and simple embedded training that focuses on how to use the domain names of phishing and the legitimate webpages to identify a phishing webpage can be retained and transferred to other webpages even one week later.
- Using browser warnings as a medium enables cybersecurity training at scale to reach the general user population.

REFERENCES

Al-Daeef, M. M., Basir, N., & Saudi, M. M. (2017). Security awareness training: A review. In *Proceedings of the World Congress on Engineering* (pp. 446–451). London, U.K.

Anandpara, V., Dingman, A., Jakobsson, M., Liu, D., & Roinestad, H. (2007). Phishing IQ tests measure fear, not ability. In S. Dietrich & R. Dhamija (Eds.), *Financial cryptography and data security*. (Lecture Notes in Computer Science, vol. 4886, pp. 362–366). Berlin: Springer.

Anderson, J. R. (2013). *The architecture of cognition*. New York, NY: Psychology Press.

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

Canfield, C. I., Fischhoff, B., & Davis, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. *Human Factors*, *58*, 1158–1172.

Caputo, D. D., Pfleeger, S. L., Freeman, J. D., & Johnson, M. E. (2014). Going spear phishing: Exploring embedded training and awareness. *IEEE Security & Privacy*, *12*, 28–38.

Carpenter, S., Zhu, F., & Kolimi, S. (2014). Reducing online identity disclosure using warnings. *Applied Ergonomics*, *45*, 1337–1342.

Chou, N., Ledesma, R., Teraguchi, Y., & Mitchell, J. C. (2004). Client-side defense against web-based identity theft. In *Proceedings of the 11th Annual Network and Distributed System Security Symposium*. http://crypto.stanford.edu/SpoofGuard/webspoof.pdf

Dhamija, R., & Tygar, J. D. (2005). The battle against phishing: Dynamic security skins. In *Proceedings of the 2005 Symposium on Usable Privacy and Security* (pp. 77–88). New York, NY: ACM.

Dhamija, R., Tygar, J. D., & Hearst, M. A. (2006). Why phishing works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 581–590). New York, NY: ACM.

Dodge, R. C., Carver, C., & Ferguson, A. J. (2007). Phishing for user security awareness. *Computers & Security*, *26*, 73–80.

Downs, J. S., Barbagallo, D., & Acquisti, A. (2015). Predictors of risky decisions: Improving judgment and decision making based on evidence from phishing attacks. In E. A. Wilhelms & V. F. Reyna (Eds.), *Neuroeconomics, judgment, and decision making* (pp. 239–253). New York, NY: Psychology Press.

Egelman, S., Cranor, L. F., & Hong, J. (2008). You've been warned: An empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1065–1074). New York., NY: ACM.

FBI (2018). 2017 *Internet crime report*. Retrieved from https://pdf.ic3.gov/2017_IC3Report.pdf

Felt, A. P., Ainslie, A., Reeder, R. W., Consolvo, S., Thyagaraja, S., Bettes, A., & Grimes, J. (2015). Improving SSL warnings: Comprehension and adherence. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 2893–2902). New York, NY: ACM.

Ferguson, A. J. (2005). Fostering e-mail security awareness: The West Point carronade. *Educause Quarterly*, *28*, 54–57.

Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to detect phishing emails. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 649–656). New York, NY: ACM.

Fu, A. Y., Liu, W. Y., & Deng, X. (2006). Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD). *IEEE Transactions on Dependable and Secure Computing*, *3*, 301–311.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of *d'*. *Behavior Research Methods, Instruments, & Computers*, *27*, 46–51.

Healy, A. F., & Bourne, L. E., Jr. (Eds.) (2012). *Training cognition: Optimizing efficiency, durability, and generalizability*. New York, NY: Psychology Press.

Herzberg, A., & Gbara, A. (2004). Trustbar: Protecting (even naive) web users from spoofing and phishing attacks. *Cryptology ePrint Archive*, Report 2004/155. http://eprint.iacr.org/2004/155.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC). A framework for perception and action. *Behavioral & Brain Sciences*, *24*, 849–937.

Hong, J. (2012). The state of phishing attacks. *Communications of the ACM*, *55*, 74–81.

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, *50*, 94–100.

Kelley, C. M., Hong, K. W., Mayhorn, C. B., & Murphy-Hill, E. (2012). Something smells phishy: Exploring definitions, consequences, and reactions to phishing. In *Proceedings of the 56th Human Factors and Ergonomics Society Annual Meeting* (pp. 2108–2112). Santa Monica, CA: Human Factors and Ergonomics Society.

Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing detection: A literature survey. *IEEE Communications Surveys & Tutorials*, *15*, 2091–2121.

Kumaraguru, P., Cranshaw, J., Acquisti, R., Cranor, L., Hong, J., Blair, M. A., & Pham, T. (2009). A real-word evaluation of anti-phishing training (Technical report) Pittsburgh, PA: Carnegie Mellon University.

Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L. F., Hong, J., & Nunge, E. (2007). Protecting people from phishing: The design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 905–914). New York, NY: ACM.

Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., & Hong, J. (2010). Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology*, *10* (2), Article 7.

Laughery, K. R., & Wogalter, M. S. (2006). Designing effective warnings. *Reviews of Human Factors and Ergonomics*, *2*, 241-271.

Lee, Y. S., & Vakoch, D. A. (1996). Transfer and retention of implicit and explicit learning. *British Journal of Psychology*, *87*, 637–651.

Lin, E., Greenberg, S., Trotter, E., Ma, D., & Aycock, J. (2011). Does domain highlighting help people identify phishing sites? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2075–2084). New York, NY: ACM.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163–203.

Oberauer, K. (2010). Declarative and procedural working memory: Common principles, common capacity limits? *Psychologica Belgica*, *50*, 3–4.

Orgill, G. L., Romney, G. W., Bailey, M. G., & Orgill, P. M. (2004). The urgency for effective user privacy-education to counter social engineering attacks on secure computer systems. In *Proceedings of the 5th Conference on Information Technology Education* (pp. 177–181). New York, NY: ACM.

Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., & Jerram, C. (2015). The design of phishing studies: Challenges for researchers, *Computers & Security*, *52*, 194–206.

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In T. N. Postlethwaite & T. Husen (Eds.), *International encyclopedia of education* (2nd ed.; pp. 6452–6457). Oxford, England: Pergamon Press.

PhishTank. (2018). Stats. Retrieved from https://www.phishtank.com/stats.php

Prakash, P., Kumar, M., Kompella, R. R., & Gupta, M. (2010). Phishnet: Predictive blacklisting to detect phishing attacks. In *Proceedings of INFOCOM, IEEE* (pp. 1–5). Piscataway, NJ: IEEE.

Proctor, R. W., & Chen, J. (2015). The role of human factors/ergonomics in the science of security: Decision making and action selection in cyberspace. *Human Factors*, *57*, 721–727.

Roediger III, H. L., Dudai, Y., & Fitzpatrick, S. M. (Eds.) (2007). *Science of memory: Concepts*. New York, NY: Oxford University Press.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217.

Sheng, S., Wardman, B., Warner, G., Cranor, L. F., Hong, J., & Zhang, C (2009, July). An empirical analysis of phishing blacklists. In *Proceedings of the 6th Conference on Email and Anti-Spam, CEAS'09*. Mountain View, CA.

Stone, A. (2007). Natural-language processing for intrusion detection. *Computer*, *40*, 103–105.

Whittaker, C., Ryner, B., & Nazif, M. (2010). Large-scale automatic classification of phishing pages. In *Proceedings of the Network and Distributed System Security Symposium, NDSS 2010*, San Diego, CA.

Wogalter, M. S., Godfrey, S. S., Fontenelle, G. A., Desaulniers, D. R., Rothstein, P. R., & Laughery, K. R. (1987). Effectiveness of warnings. *Human Factors*, *29*, 599–612.

Wogalter, M. S., & Mayhorn, C. B. (2008). Trusting the internet: Cues affecting perceived credibility. *International Journal of Technology and Human Interaction*, *4*, 75–93.

Wu, M., Miller, R. C., & Garfinkel, S. L. (2006). Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 601–610). New York, NY: ACM.

Xiong, A., Proctor, R. W., Yang, W., & Li, N. (2017). Is domain highlighting actually helpful in identifying phishing web pages? *Human Factors*, *59*, 640–660.

Yang, W., Xiong, A., Chen, J., Proctor, R. W., & Li, N. (2017). Use of phishing training to improve security warning compliance: Evidence from a field experiment. In *Proceedings of the Hot Topics in Science of Security: Symposium and Bootcamp* (pp. 52–61). New York, NY: ACM.

Zhang, Y., Hong, J. I., & Cranor, L. F. (2007). Cantina: A content-based approach to detecting phishing web sites. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 639–648). New York, NY: ACM.

Aiping Xiong is an assistant professor in the College of Information Sciences and Technology at the Pennsylvania State University in University Park. She earned her MS in industrial engineering in 2014 and PhD in Cognitive Psychology in 2017 from Purdue University.

Robert W. Proctor is a distinguished professor in the Department of Psychological Sciences at Purdue University, West Lafayette, Indiana. He received his PhD in experimental psychology from the University of Texas at Arlington in 1975.

Weining Yang works at Google, Inc. He received his PhD in computer science from Purdue University in August, 2016.

Ninghui Li is a professor in the Computer Science Department at Purdue University. He received his PhD in computer science from New York University in 2000.

Date received: June 2, 2017
Date accepted: October 7, 2018