Use of Phishing Training to Improve Security Warning Compliance: Evidence from a Field Experiment

Weining Yang[1], Aiping Xiong[1], Jing Chen[2], Robert W. Proctor[1], Ninghui Li[1]
[1]Purdue University [2]New Mexico State University
{yang469, xionga, rproctor, ninghui}@purdue.edu jingchen@nmsu.edu

ABSTRACT

The current approach to protect users from phishing attacks is to display a warning when the webpage is considered suspicious. We hypothesize that users are capable of making correct informed decisions when the warning also conveys the reasons why it is displayed. We chose to use traffic rankings of domains, which can be easily described to users, as a warning trigger and evaluated the effect of the phishing warning message and phishing training. The evaluation was conducted in a field experiment. We found that knowledge gained from the training enhances the effectiveness of phishing warnings, as the number of participants being phished was reduced. However, the knowledge by itself was not sufficient to provide phishing protection. We suggest that integrating training in the warning interface, involving traffic ranking in phishing detection, and explaining why warnings are generated will improve current phishing defense.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; K.4.4 [**Electronic Commerce**]: Security

Keywords

Phishing; Field Study; Active Warning

1. INTRODUCTION

Phishing is a widespread and continuously evolving threat in cyber security [2]. Usually, phishing campaigns start from spoofed emails or instant messages that contain links directing users to "fake" sites, where they are asked to provide personal information. As the fake sites often look identical or similar to legitimate ones, users are tricked into entering sensitive information, which is then stolen by the attackers.

The current approach to protect users from phishing attacks is to display a warning when the webpage is considered suspicious, and ask users to decide whether to visit the webpage. If phishing detection could be done with 100% accuracy, then a phishing defense mechanism could unilaterally decide to block a page identified as phishing, just as today's anti-virus software makes quarantine decisions without user inputs. But in phishing detection, as in the case of many other computer security defenses, the correct decisions are often context-dependent, and it is impossible for a detection mechanism to achieve 100% accuracy. Additionally, there is a tradeoff between false positive and false negative. To ensure that warnings are shown for the vast majority of phishing pages, some warnings are incorrect. As a result, having humans in the loop to make the decisions is unavoidable [13, 17].

When human decisions are needed, understanding of the warning becomes an important factor. Current warning designs focus on describing the potentially dangerous outcome if the warning is not heeded. Existing literatures on improving comprehension recommend avoiding technical jargon, using simple language [7, 19, 38], being as brief as possible, and describing the specific risks clearly [20]. One aspect that has received little attention is explaining the rationale behind the warning.

We hypothesized that when users are able to understand why a warning message is generated, they are capable of making correct decisions. We evaluated this hypothesis with a field study on phishing warnings. To perform such a study, we needed a warning signal that could be easily explained to users. We observed that the vast majority of phishing pages are hosted on new domains registered for the purpose of malicious attacks or infrequently used websites. At the same time, phishing pages try to masquerade as popular sites (such as Amazon, eBay, or Taobao). We expected that highlighting this domain ranking mismatch between the expected authentic websites and phishing websites is easy for users to understand and likely to catch their attention.

Most phishing warning research has been conducted as role-playing experiments in laboratory settings [16, 39]. However, the extent to which data obtained from showing phishing warnings in a lab setting can reflect users' behavior in the actual field setting is questionable. There are many reasons why participants' actions when encountering a phishing warning in a lab setting may be different from what they would do in a natural setting [30, 34]. Thus, we conducted a controlled field experiment in which we "phished" for participants' Amazon passwords during their daily computer use. Our simulated phishing attack is able to bypass today's phishing defense and reach the participants.

The main contribution of the paper is as follows:

- We report results of a phishing defense study in the

HoTSoS, April 04-05, 2017, Hanover, MD, USA

DOI: http://dx.doi.org/10.1145/3055305.3055310

user's natural everyday computer use, which is critical in understanding the effectiveness of phishing defenses.

- Our findings indicate an interaction between training and the effectiveness of warning messages. Phishing training greatly enhanced the effectiveness of phishing warnings. We observed 50% click-through rate for participants who did not receive training, and 0% click-through rate for participants who were equipped with procedural knowledge about phishing. However, training by itself was insufficient in defending against phishing. Without a warning, almost all participants fell for our simulated phishing attack.

- Our findings suggest several improvements to today's phishing defense, including the following: (1) integrating training of phishing into the user experience; (2) weighting traffic ranking more heavily in the overall phishing detection architecture; and (3) explaining why phishing warnings are generated.

The paper is organized as follows. We start with reviewing related work on phishing webpage detection, warning interface design and procedural knowledge, followed by presenting an overview of the study. Then, we describe a pilot study and the main study. After that, we discuss a) potential ways to improve the current phishing defense using the findings from the study, b) other observations, and c) limits of our study.

2. RELATED WORK

In phishing defense based on blacklist, URLs are checked against predefined blacklist(s) of known phishing sites. The effectiveness of blacklist-based methods has been widely studied [27, 33, 41], and a high false negative rate (i.e., phishing URLs not detected) is a common issue [41]. A good blacklist can cover more than 90% of the live phishing URLs [27], but the detection rate for a new phishing site may be less than 20% [33]. Thus, blacklists are ineffective at protecting users during the initial phase of phishing attacks.

Machine learning methods have also been extensively applied in phishing detection. Such methods classify a web page's legitimacy dynamically by using several features extracted from URLs and/or page content. Representative approaches include Xiang et al. [40] and Ma et al. [28, 29]. Whittaker et al. [35] developed a system that uses most features considered in the literature to maintain Google's phishing blacklist. While these systems can get quite high accuracy rates, they are generally heavy-weight and cannot be deployed on the browser end. A web browser must deploy a simpler and less accurate detection mechanism, and send URLs that are suspicious to the servers for more detail analysis.

Passive phishing warnings indicate potential dangers to the user without interrupting a user's primary task, for example, changing toolbar color [10], providing textual information [22] and dynamic security skins [14], or highlighting the domain name in the address bar [26]. Wu et al. [39] investigated three anti-phishing toolbars intended to help users determine when they are interacting with fraudulent websites. Most participants ignored passive security indicators, relying instead mainly on the web site's content to decide its trustworthiness. The participants who noticed the warnings assumed that the warnings were invalid because they did not understand the warnings.

Active warnings [5, 19] were shown to capture a user's attention by forcing the user to choose one of the options presented by the warnings. Although the adherence rate increased compared to passive warnings, a substantial click-through rate was still evident for active warnings [18, 19]. A common issue reported across those active warning studies was that users often did not understand the warning's meaning. According to the Communication-Human Information Processing (C-HIP) model [37], warnings are processed through attention, comprehension, and action sequentially within a certain environment. Thus, an ideal compliance rate cannot be achieved if warning comprehension is poor. To improve users' comprehension, prior warning interfaces, like that proposed by Felt et al. [19], were designed based on best practices from the warning literature. Steps such as avoiding technical jargon, being brief, targeting low reading level, and describing specific consequences, were taken to communicate the possible risks.

Based on C-HIP, Cranor [13] introduced a human-in-the-loop framework in a secure system by adding users' personal variables, intentions and capabilities. Within this framework, the users' general knowledge of cybersecurity or phishing is closely related to their understanding of the warning and their final actions. Downs et al. [15] investigated differences between declarative knowledge (verbalizable facts) and procedural knowledge (actions to take) in an online role-play study regarding possibly fraudulent emails and possible actions for webpages following links in the emails. Although declarative knowledge (understanding the security terms) was closely related to the participants' self-reported predictions on awareness, susceptibility and intentions, it was not related to changes in behaviors. In contrast, procedural knowledge (determining URL legitimacy) was the only predictor for the users' ability to adjust their risk decisions. Thus, the lack of procedural knowledge probably explains the remaining click-throughs even when the safe choice was promoted by the preferred option design in [19].

As noted, most phishing warning research has been conducted as role-playing experiments in laboratory settings [15, 16, 39]. However, participants' awareness of their participation in a phishing study influences their actions [30, 34]. To our knowledge, the only field study about the effectiveness of phishing warnings was conducted by Akhawe et al. [5], in which adherence rates of browser security warnings in Google Chrome and Mozilla Firefox were compared. The data included over 25 million warning impressions collected by telemetry frameworks from users' normal browsing activities. Users continued through less than a tenth of malware and phishing warnings of Mozilla Firefox, and a quarter of warnings in Google Chrome, indicating that active browser security warnings can be effective in practice.

3. STUDY OVERVIEW

3.1 *D*omain *R*ank based *A*nti-phishing *W*arning (*DRAW*)

For the purpose of understanding the efficacy of warning messages, the effectiveness of training, and the interaction between the two, we needed a warning signal that could be easily explained to users. We chose to use the traffic ranking of domains as the warning criteria. This is motivated

by the observations that the targeted websites of phishing attacks are mostly popular ones, and at the same time, most phishing pages are hosted on infrequently visited sites. We designed *DRAW*, a *D*omain *R*ank based *A*nti-phishing *W*arning, which alerts users when phishing conditions are likely to be satisfied. More specifically, when a user attempts to enter information on a webpage hosted on a domain that has traffic ranking greater than a threshold (set to 100,000 in our main study), *DRAW* displays a warning dialog. The warning highlights the domain name as well as the fact that it is infrequently visited by users. No warning is shown if the user does not enter any information, or if the domain belongs to a local whitelist created based on the browsing history. *DRAW* is implemented as a Chrome extension and obtains the traffic rankings of domains from Alexa.com.

We want to emphasize that *DRAW* is not a complete phishing defense system. It is a vehicle for studying the interaction between warning understanding and effectiveness. However, the main idea of *DRAW* can be integrated into today's phishing defense system. For example, instead of raising a warning as *DRAW* does, the web browser can send the URL to the server for more in-depth analysis. Traffic ranking is a difficult feature for attackers to manipulate when they attempt to evade detection. The cost of using a high-ranking domain for phishing, by compromising a popular website or setting up a website and boosting its popularity, is high. This cost is because a high-ranking domain is typically owned by parties who have incentives to protect it and root out phishing pages once the domain is exploited. While an attacker can own a domain and spend resources to improve its traffic ranking, once the domain is detected, it can be blacklisted. Since most phishing pages can be up for only a few hours before they are reported and blocked, successful blocking of phishing attacks from newly created domains or obscure domains will significantly raise the bar for attackers to carry out phishing attacks.

3.2 Application of a Four-step Iterative Process Framework

Our experiment design followed a four-step iterative human threat identification and mitigation process proposed by Cranor et al. [13] to identify the cause of observed failures and find ways to mitigate them. The four steps are 1) *Task identification*: Having humans involved in making the decisions identifying phishing webpages is unavoidable given today's detection accuracy. 2) *Task automation*: We provided the rationale of popping up the warning, and the potential outcome of visiting the page in a concrete but easy-to-understand manner. The warning is expected to help users make an informed decision balancing risk and usability. 3) *Failure identification*: We designed a pilot user study to observe people's reactions to phishing emails and phishing sites in the wild, and to evaluate the efficacy of *DRAW*, and identified some factors causing failures. Most participants fell for the simulated phishing attack due to (a) a failure to understand the warning because of the interface design; and (b) the low risk of the account targeted in the study. 4) *Failure mitigation*: We thus redesigned the user study accordingly. In the main study, we switched the phishing target to Amazon and improved the warning interface.

4. PILOT STUDY

We conducted a pilot study, approved by the our institutional review board (IRB), to assess how frequently users will visit and enter information into the websites for which *DRAW* will raise a warning, to test our warning interface design and the study protocol, and to receive feedback from participants.

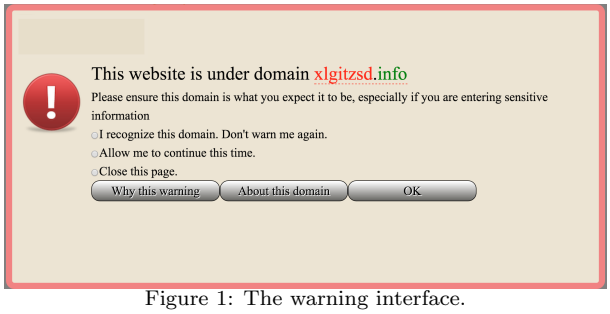

Figure 1: The warning interface.

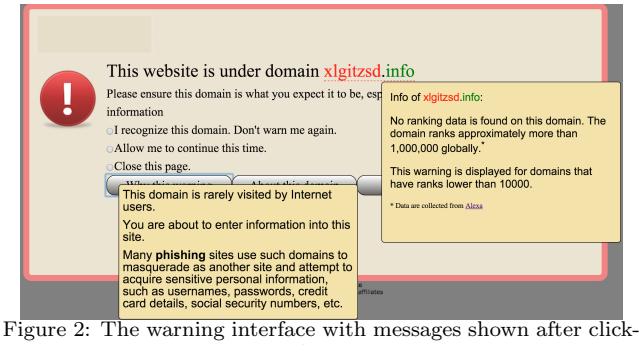

Figure 2: The warning interface with messages shown after clicking "Why this warning" and "About this domain".

4.1 Warning Interface Design

The design of the warning interface used in the pilot study is illustrated in Fig 1 and Fig 2. We extracted the domain name from a URL and marked it in different colors in order to aid users in determining the current webpage's legitimacy. In the warning, three options are listed: (1) *I recognize this domain. Don't warn me again*; (2) *Allow me to continue this time*; (3) *Close this page.* The first option adds the domain to a local whitelist so that warnings will not pop up on the same domain again. The second option allows users to visit the page only for one time. When the user tries to enter information on the same domain in another visit, the warning will be generated again. The third option is to close the webpage. Users can also get more detailed descriptions about phishing by clicking the "Why this warning" button, and the detailed ranking information of the domain if the "About this domain" button is clicked.

4.2 Procedure

The pilot study lasted for 6 weeks, and 9 participants were involved. The major component of the study was a simulated phishing attack conducted by sending forged survey emails, which is a procedure that has been deployed to induce people into entering their account information [11]. However, this component was concealed to participants, in order to preserve a natural setting and not to influence participates' behavior. Instead, we named the study "Internet browsing behavior study", and told participants that the purpose of the study was to understand people's behavior on Internet browsing. We set up a website that allowed par-

ticipants to register for a new account, log in, and complete surveys we provided. The participants were recruited by way of fliers placed around the campus.

At the beginning of the study, the participants came to our lab to install the Chrome extension. The participants were presented with the cover story and asked to create new accounts on our survey website with an email address and a password. The participants were told that each week they were required to login to our website and finish a short questionnaire that asked for an estimation of the participant's time allocation on different online activities. At no time during the study, did we mention phishing or hint at phishing protection. Each week, we sent an email to all participants asking them to finish the questionnaire. In the email, a shortened URL of our survey site (using tiny URL[1]) was included, to hide the domain of the URL in the email. In weeks 1-5, the links in the emails directed the participants to the legitimate website where they registered. In the last week of the study, the links in the emails were associated with a newly registered domain maintained by us, to simulate a phishing attack. The content of this website was identical to the normal survey site except that the URL was different. We considered a participant to be "phished" if correct account information was entered on the "fake" website.

Participants were divided into two groups: control group (3 participants), in which no warning was presented during the whole experiment time, and experimental group (6 participants), in which *DRAW* functioned normally. Participants in the experimental group saw the warning described previously when they attempted to enter information on domains ranked greater than 10,000, including the "phishing" domain used in the study. The warning threshold (10,000) was set to ensure that participants were likely to have seen our warning a few times before the simulated phishing. At the end of the study, the participants came to our lab again for a semi-structured interview, after which they were debriefed about the true purpose of the study and informed about the hypothesis for the experiment as well.

4.3 General Website and Warning Statistics

During the 6-week study period, each participant visited 7,985.3 webpages on average with a standard error (SE) of 3,302.5. The average number of unique domains visited was 340.7 with a SE of 47.6. The mean number of web pages with information entered was 573.1, and the SE is 83.9, and the average number of unique domains where information was input was 61.0 with a SE of 8.3. Excluding warnings displayed on our "phishing" site, the average number of warnings generated for each participant (not including the control group) was 14.7 with SE of 3.9, and the maximum number was 28. On average, a warning would be generated once for every 543.2 webpages a participant visited, or once for every 40.9 webpages where the participant entered information. On average, the option "trust the domain" was chosen 11.6 times. The warning frequency was greater than we had expected. In fact, 5 participants mentioned seeing warning on regularly visited sites, and 1 participant complained about the high frequency of warnings in the post-session interview.

4.4 Simulated Phishing Result

During the "phishing" week, all 3 participants in the control group entered their passwords and attempted to login, and the passwords from 2 of them matched the genuine passwords we recorded. Unexpectedly, among the 6 participants in the experimental group, none of them chose to "close the page", 5 chose to permanently trust our "phishing" domain, and 1 chose to "continue this time". However, only 1 participant (who chose to permanently trust our domain) entered the correct password. To understand the rationale behind the participants' behaviors, we interviewed them regarding phishing and the warning generated by *DRAW* at the end of the study. It turned out that only one participant in the experimental group, although she chose to permanently trust our domain, intentionally provided a wrong password, while the rest entered mismatched passwords by accident. We found that participants chose to enter genuine information because of two reasons: (1) they did not understand the meaning of our warning and tended to ignore it in part due to the interface design; (2) they believed that the risk of the accounts being stolen is low because nobody other than our research group knew the survey websites or would interacted with the accounts on the site. Another surprising finding was that almost half (4/9) of the participants had not heard about or did not know the meaning of phishing.

[1]http://tinyurl.com

5. MAIN STUDY

5.1 Experiment Design

Findings from the pilot study guided us to redesign the warning interface and the study scenario, and to add another factor, phishing training. The main study differs from the pilot study in the following aspects:

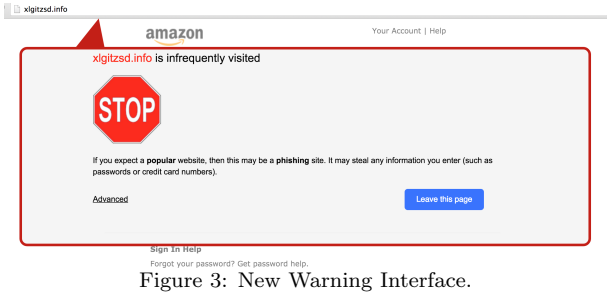

Figure 3: New Warning Interface.

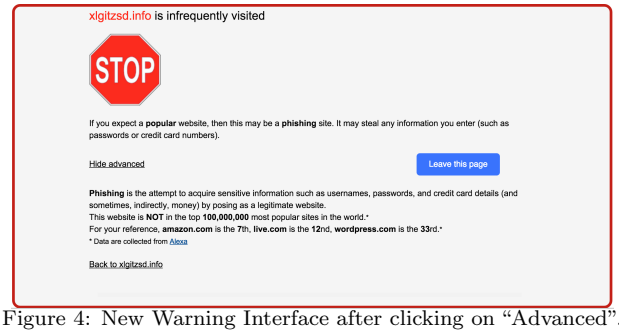

Figure 4: New Warning Interface after clicking on "Advanced".

Improving comprehension of the warning.

In the main study, we redesigned the warning interface in order to produce a better comprehension from the aspects of both declarative knowledge and procedure knowledge. The updated warning interface is illustrated in Fig 3 and Fig 4. Since *DRAW* is installed in Google Chrome, we kept the

layout of our warning consistent with the current built-in Chrome SSL/Phishing warning [19]. We did not adopt the red color theme of the built-in Chrome Phishing warning, because red is generally associated with danger and stop, but we wanted to encourage participants to go ahead and make their own judgments based on the information provided in the warning. Instead, we selected gray, which was suggested to have a high adherence rate by Felt et al. [19]. Besides the background color, a fundamental difference between our warnings and Chrome's built-in warnings was the extra designs to provide extra information to users in order to help them make informed decisions. Instead of simply telling general facts or declarative knowledge regarding phishing, we provided more detailed and site-specific cues to help users detect potential deceptions. On the top of the warning, the domain name was extracted from the URL as in the pilot study. But the domain name was enlarged, marked in red, and linked to the URL by an arrow, indicating that this warning was specific for the domain of current webpage. In addition, we described the current domain as "*is infrequently visited*". Thus, whenever users expect to visit a popular legitimate website, such as Amazon, the mismatched domain name and visit frequency are provided as cues to the users in order to help them understand why the warning is presented. Also, to provide more context for the traffic ranking numbers, three popular websites' rankings were listed below the current webpage's ranking. Those 3 websites were randomly selected from the first 100 most frequently visited websites in the United States.

We redesigned the presentation of the declarative knowledge as well, according to the guideline of efficient warning proposed by Laughery et al. [25], in which three key components are included: (1) user's subjective understanding of the risk connotation; (2) the direct language and symbol; (3) user's background knowledge. To address the first aspect, signal words are often included within a warning to indicate the level of the risk present, which has been shown to increase user's risk perception and the warning effectiveness. For the second perspective, specific, complete, but not too lengthy text to describe the nature of the risk as well as its explicit consequences enhances the communication of the warning is recommended. Finally, simple language should be used to target as many users as possible. According to the guidelines, we updated the design of the warning dialog with the following changes: (1) A "STOP" sign was implemented to convey the message to the user that they need to stop and check. (2) The content of the warning was described in a specific and complete way, including the information of phishing, instructions on how to avoid it, and the explicit potential consequence if it is not avoided. (3) The description was not so lengthy that few people would take time and effort to read it. (4) Taking the background knowledge of general online users into account, simple texts were used to keep the reading levels as low as feasible, which also excluded any highly technical information or jargon words [19, 37].

In addition, directive action (i.e., opinionated action) [19] was provided in our warning, which is the "Leave this page" button highlighted in blue. Opportunity to go to the infrequently visited webpage was also provided by the "Advanced" button. Instead of directly going back to the risky webpage, extra steps were added to make it more difficult for users to proceed past the warning [5]. A detailed explanation of phishing was first listed, followed by the detailed ranking information of current page. A "Back to XYZ.com" button was listed in the end if the user would still like to go to the webpage.

Targeting more sensitive accounts. We decided to use Amazon as the target in the simulated phishing attacks because it is widely used among students and eCommerce takes the largest part of phishing attacks [1]. By doing that, the target in our phishing simulation is no longer a newly created account, which might not be carefully managed by users. To simulate a real phishing attack, the overall appearance was identical to the Amazon login page except the domain name was *xlgitzd.info* and links such as *Forgot your password?* or *Help* were disabled. The fake Amazon login page was maintained by us.

Figure 5: Email that spoofs an Amazon gift card

We copied common phishing emails that spoof Amazon gift card and changed the content to fit our study. The message claimed to be from Amazon, and was sent 3 times. In the first 2 times, the sender was set as *gc-orders@gc.email.amazon.com*, which is the origin of real Amazon gift cards. However, some major email providers (e.g., Gmail) block the email due to Email Authentication, which ensures that an email provider will be able to recognize the sender of an incoming message and verify the source of messages received [3]. To get around such protection and ensure that the participants received our phishing emails, in our third phishing email we set the sender as *gc-orders@www-258389988.us-east-1.elb.amazonaws.com*, which is a random email address. Our phishing email can be found in Figure 5. To prevent users' checking the authenticity of the gift card, we covered the gift card number with stars except for the last four digits. Since participants no longer needed a long study period to get familiar with the targeted site (Amazon), we reduced the length of the study to 3 weeks.

Other changes. We noticed that the warning rate in the pilot study was too high. In the main study, we changed the the warning threshold to $100,000$. We further disabled the warning on webpages stored as bookmarks. Observing the lack of phishing knowledge, in the main study we introduced a short training session regarding general knowledge about phishing at the beginning of the study for some participants, and evaluated the effect of the training.

5.2 Methodology

Recruitment and Demographics. Participants were recruited in the same way as in the pilot study, i.e., by fliers placed around the campus, and only Chrome browser users were invited. Again, the study was presented to participants as an "Internet browsing behavior study". 63 participants (34 males) were recruited with 84% (49) of them aged from 19 to 22 years old, and the remaining ones were all less than 30 years old except one. 61% (58) of the participants were undergraduate students, and the rest had higher degrees. The study was approved by IRB as well.

Procedure. Those who replied to our advertisement were invited into our lab for a short starting session. Each participant was assigned randomly to one of two groups, which were identical except that one group included an additional short phishing training. For both groups, experiment details and *DRAW* installation instructions were presented in the same slides, and extra training slides were presented near the end for the training group only. In the group with phishing training, the definition of phishing was provided and a banking phishing email example was presented. Participants were also taught how to evaluate the legitimacy of a URL by identifying the domain name. In addition, participants were tested with a list of URLs that included both legitimate and fraudulent types, with feedback provided. We also mentioned that *DRAW* might help protect them from phishing attacks without revealing more details. The training lasted approximately 10 minutes. With participants' agreement, *DRAW* was then installed on their own laptops to record their interactions with the browser, including hash values and traffic rankings (only one significant digit was record) of the domains visited, and all the interactions with *DRAW*, for the duration of the study. We told participants that they might have some interactions with our browser extension without going into details (and nobody asked about the details). At the end of the session, all participants were told that there would be a lottery at the end of study, and the winner would be granted an Amazon gift card (the amount was not mentioned) sent by email.

Once *DRAW* was installed, it first loaded all bookmarks in the browser and added all the domains of URLs in bookmarks into a local whitelist. To further reduce warning rates, in the first week of the study, *DRAW* was in a "training" mode and and did not generate any warning. Instead, it added all visited domains during the "training" period into the whitelist as well. After the first week, *DRAW* functioned normally. Once a user attempted to enter information on a webpage with the domain whose ranking exceed the threshold $(100,000)$ and the domain was not in the whitelist, a warning dialog would pop up. If the user ignored the warning and chose to visit the page anyway, the domain would be added into the whitelist.

In the middle of the third week, we sent fake Amazon gift card emails described above. Participants were directed to our fake Amazon webpage. If a participant attempted to enter information on the page, the phishing warning was triggered. For those participants who did not visit the fake Amazon login page, we kept sending the same phishing email for another two days. In the last email, the sender was changed to the random email address. The participants who visited the phishing webpage but did not enter any information, or those who ignored the phishing email for 3 times, were identified as not falling for phishing. Participants who entered both account email address and password in the spoof Amazon login page, were redirected to an online survey to verify whether the information input was genuine. We did not record the participant's input on the fake login webpage. Instead, we only retrieved (but did not record) the length of the passwords in the browser and asked questions used by [9]: "You entered l characters into the password field of the previous webpage. This password you entered is stored in your browser but we have not sent your password to our server. Was the password you entered a real password for your account in Amazon?" If the answer for the first question was no, a follow-up question is shown: "Since you did not enter a genuine password into the webpage, may we collect the content of this field for analysis?". If a participant answered "yes" in the first question, or answered "no" in the second one, we considered the participant to have fallen for phishing.

At the end of the study, all participants were required to finish a post-session questionnaire which included demographic questions and questions regarding the frequency of our warning and participants' reactions to the warning. Upon the payment for the study, each participant was provided with a debriefing form that discussed the objectives and methods of the study.

5.3 General Websites and Warnings Statistics

During the 3-week study period, each participant visited 3653.5 webpages on average with a SE of 308.5. The average number of unique domains visited was 211.2 with a SE of 18.3. The mean number of web pages with information entered was 395.0, and the SE is 64.0. The average number of unique domains where information was input was 35.3 with a SE of 2.2. Note that the difference of values comparing with that of the pilot study is mainly due to the reduced length of the study. Most of the pages visited and pages on which users input information are hosted on domains with ranking not larger than 100,000. These pages contributed to 97.9% of pages that participants visited and 99.3% of pages where information was input, respectively.

The average number of warnings generated for each participant (not including the control group) was 2.9 with a SE of 0.33. Warnings triggered on our phishing page are not counted. A warning would be generated on every 1268.8 webpages one participant visited or on every 137.2 webpages on which one participant entered information, which is significantly less that in the pilot study, due to the raised warning threshold and the initialization of the whitelist with bookmarked pages and pages visited during the first week.

5.4 Experiment Results and Analysis

Identification of the phishing without warning. Among the 63 participants recruited, 30 were given the phishing training while the other 33 were not. In the simulated phishing attack, 46 participants attempted to enter information on our phishing website.

Table 1 lists the number of participants who visited our phishing site, who identified phishing campaign before entering information, and who started entering information on the page. The identification of phishing before entering information was determined from the post-session questionnaires. 4 participants with training provided and 2 participants without receiving training recognized the phishing

Table 1: Number of participants who visited our phishing page, entered information, and fell in the attack by group condition. Pwd stands for password.

Training	Total	Identified Phishing Email	Visited Phishing Page	Identified Phishing Page	Warning	Total	Submit Form	Input Genuine Pwd
Yes	30	4	24	4	Yes	12	0	0
					No	8	8	8
No	33	2	27	0	Yes	14	7	7
					No	12	12	12

email and did not go to our "phishing" site. The remaining 6 participants who did not visit our "phishing" site either did not see the email, or wanted to redeem the gift card later. For those who visited our "phishing" webpage, 5 participants did not enter information on the page. Among these participants, 4 from the group with training all identified the phishing page, where 3 of them identified the page by domain, and 1 did not see password auto-filled. The only participant who did not receive phishing training and did not enter information claimed that she visited the site on phone, and wished to login and redeem the gift card later because entering password on the phone is difficult.

Overall, 8 participants in the group with training provided and 2 participants in the group without phishing training identified the phishing campaign. The different behaviors between the two groups mainly come from the effect of phishing training, as no warning had been displayed. It appears that phishing training has some beneficial effect, although the effect is relatively small and is not statistical significant due to the sample size.

Effect of the warning and training. For the 46 participants who started entering information on the phishing page, 20 did not see warnings generated by *DRAW*, as they did not visit the page using Chrome with *DRAW* installed. We regrouped these participants based on two factors: phishing warning and phishing training. Table 1 shows the regrouping of participants and the number of participants who clicked through the warning from *DRAW* and fell in the attack. For the 27 participants who submitted the form, 26 answered "yes" to our first verification question, and the other participant answered "no" to the second verification question. Therefore, we believe all the participants who submitted the form used their genuine passwords. All participants who did not see the phishing warning (groups 2 and 4) provided their genuine account information to us (100%), regardless of whether phishing education was provided or not. In comparison, for those who saw warnings, the success rate of the phishing attack was 26.9% (7/26), which is significantly lower ($\chi^2_{(1)} = 24.900$, $p < .001$), indicating a positive effect of *DRAW* against phishing attacks.

The behavior of participants who saw warnings depended on whether or not phishing training was taken. For those equipped with both phishing warning and phishing training (group 1), no participants fell prey to the simulated phishing attack (0% successful phishing rate). In contrast, with only warning presented (group 3), half of the participants submitted genuine account information (50%). The difference in success rate of phishing attacks between the two groups is significant ($\chi^2_{(1)} = 8.211$, $p = .006$), indicating that phishing training helped participants understand our warning and thus, make correct decisions.

Comprehension of the warning. Table 2 shows the distribution of participants' behaviors to warnings generated by *DRAW* before visiting our "phishing" site. 4 participants consistently chose to trust the page on all webpages triggering warnings, including our "phishing" page, while 2 participants consistently chose to leave the page. The consistent choices might result from blindly trusting or ignoring all warnings. On the other hand, 13 participants who had seen warning on both our "phishing" site and other websites made different choices depending on websites visited, indicating their ability to make choices depending on different contexts, which suggested their own understanding of our warning message. Among these participants, 12 of them made the correct choice on our "phishing" site, and closed the page before entering any information. If only these 12 participants were considered as having correct comprehension, the comprehension rate can be calculated as $\frac{4+8}{1+4+0+4+8+2} = 63.2\%$. The actual comprehension rate might be higher, as not choosing different options might also result from the fact that all the page triggering warnings are true phishing or false alarms, and participants' comprehension was not demonstrated. In the study, there were 28 participants who chose to ignore the warning and went back to the webpage at least once. We asked why they ignored the warning and made the choice in the post-session questionnaire. 24 (85.7%) of those participants claimed that they could understand the warning. For the remaining participants, 2 did not read the warning and the other 2 said that they could not understand the warning.

Table 2: Responses to warnings prior to seeing our "phishing" site. "None" means the participants did not see any warning on webpages other than our "phishing" page.

"Fake" Amazon	Responses on other warnings (Trust or leave)			
	Both	Trust only	Leave only	None
Trust	1	4	0	2
Leave	4	8	2	5

Warning Frequency. For participants who had seen the warning at least once, we asked them about the frequency of the warning in the post-session questionnaire. It turned out that 90.2% (37/41) of the participants rated the frequency as acceptable. For the remaining 4 participants, one told us that she confused our warning with the pop up by webpages, and the others did not accept the frequency mainly because they saw warnings on legitimate webpages. These results show that the frequency of our warning is reasonably low in general.

6. IMPROVING TODAY'S PHISHING DEFENSE

The fact that our simulated phishing attack was able to bypass today's phishing defense, and the findings from the studies, suggest several improvements to today's phishing defense. Perhaps the most widely used phishing protection

mechanism nowadays is Safe Browsing [4] in Chrome provided by Google. In Safe Browsing, phishing webpages are identified by using a combination of client-side and server-side detection. A Chrome browser downloads a blacklist of hashes of URLs that are considered to be phishing pages to the client side. When the user visits a site through the browser, it checks whether the hash of the URL is in the blacklist, and if so, sends the URL fingerprint to the server for confirmation that the URL is indeed a phishing page. Additionally, the browser also performs some local checking to detect whether the site is suspicious of being a phishing page. If the page is found to be suspicious, information of the page is sent to the server, which applies machine learning models obtained from extensive training to the page. In either case, if the server confirms that the URL is phishing, the browser displays a phishing warning page.

Our webpage for simulating a phishing attack targeting Amazon is able to bypass Chrome's browser side screening. However, we believe that it probably can be detected as a phishing page by Safe Browsing's server-side detection engine, assuming that it renders the webpage and extract visual features from it, since our page looks almost identical to Amazon's login page.

Explain the underlying reason of showing the warning. Our findings imply that combining explaining the warning criteria with training is effective. Displaying such reason for system's decision might be hard to apply on real world detections due to sophisticated machine learning models. Nevertheless, assuming the detection process identifies the targeted website of the phishing campaign, e.g., by text mining or image classification, it is feasible to display a similar message that presents the comparison between the targeted website that is expected by the user, and the real (phishing) domain. For example, a warning on our "phishing" site might be *This website, under the domain* ***xlgitzsd.info****, appears to pretend to be Amazon.*

Provide training opportunity. While our findings demonstrate the effect of phishing training, the method of using such training lecture or notes does not scale. However, it is possible to provide integrated phishing training opportunity as part of a holistic phishing defense mechanism. If a user visits a phishing site that is known for certain to be a phishing site, and the user ignores the warning page to continue, the browser can activate a second layer of defense and intercept when information is about to be sent from the browser to the website. At this point, the browser can conduct a training session, telling users that they have most likely fallen for a phishing attack, while at the same time explaining general knowledge about phishing and procedural knowledge on phishing detection. We conjecture that showing the verified facts rather than risk probability at the time that users have fallen for phishing would catch users' attention and increase the effectiveness of training.

Using traffic ranking as a deterministic factor on client-side detection. No phishing warning was raised on our "phishing" site; this suggests a space of improvement on phishing detection in Safe Browsing. We suggest adding the traffic ranking either as a standalone or as a high-weight feature when determining whether or not sending the webpage (or the footprint or features) to the server side detection.

We collected reported phishing URLs on PhishTank.com from June 6th to June 13th, 2015, and obtained 49,416 unique phishing URLs in total. Fig 6 shows the cumulative density of these phishing URLs based on traffic rankings of the domains. A point (x, y) on the curve means y percent phishing sites are hosted on domains with traffic ranking no larger than x. As there were 29,722 domains whose ranking information was not available on Alexa.com, which means the ranking of such domains were more than 1 million, the curve does not goes up to 100%. From the figure, we can observe that if we use domain traffic ranking as a standalone feature, and collect websites whose domains' traffic ranking are more than 100,000, more than 90% of phishing campaigns are covered. On the other hand, pages with low traffic rank, by definition, are infrequently visited. In our user study, these webpages only counts for 2.1% of the pages visited and 0.7% of the pages on which information is entered. Therefore, the overhead on the client side is minimal, and the overhead on the server side should be able to acceptable as well.

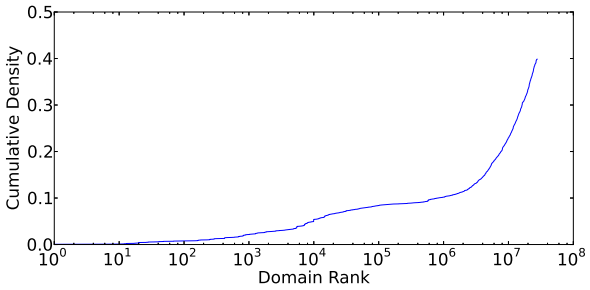

Figure 6: Cumulative Density of Phishing Sites.

7. DISCUSSION

Field Experiment. Our experiment was conducted in a field setting to mimic a real phishing scenario. This method can help us understand users' precise responses against phishing attacks in everyday computer use. It also provides a chance to validate the results of prior warning design studies. Nevertheless, our experiment also revealed the complexity and difficulty of conducting phishing scam research.

First of all, the study was time consuming, taking almost one year to conduct. Even the IRB approval process was more difficult and longer than that for surveys or lab experiments. Due to the deceptive nature of phishing, we attempted to recruit participants without strong connections with each other to avoid the possible between-subject communication. Such communication could reveal the true purpose of the study, which would harm the accuracy of the experimental result. Thus, the recruitment of participants had to be distributed across the year. Despite the length of time required to conduct the main study, the number of participants was limited to 63. Also, even though participants were requested to use Chrome as the only browser during the 3-week experiment, our results showed unusually light traffic of webpage browsing for some participants, suggesting that they used other browsers or devices. Due to such unexpected behavior, we had to increase the number of participants gradually to obtain a balance among the 4 groups.

We believe the difficulty in conducting the study is worthwhile, as the results gained from a field-setting experiment are more trustworthy than those from a laboratory exper-

iments. There are many reasons to cause participants' actions in a lab setting when encountering a phishing warning to be different from what they do in a natural setting [30, 34].

Effect of Phishing Training. Our experimental results show that the active phishing warning was effective in terms of phishing protection, which is consistent with the results from existing literatures [18, 19]. Whereas phishing training alone showed no significant advantage, a combination of phishing training and active phishing warning can significantly reduced the click-through rate.

Although previous authors [21] doubted the effectiveness of phishing education, the power of training was well demonstrated in our study. The main difference in training process between our study and the existing literature is that our training focusing on techniques regarding how to identify phishing pages, which serves as procedural knowledge, instead of only presenting facts information of phishing, which is known as declarative knowledge. And the effectiveness of training was also evident in prior studies when a training message was delivered at the time users clicked on the URL in a simulated phishing email [24].

Procedural knowledge is mainly about using knowledge to solve problems [6, 8], which can be directly applied. Procedural knowledge is often acquired implicitly through practice and repetition [31]. Once acquired, the skill appears to be unconscious [8], can be tuned towards particular types of situations [23], and applied in an automatic fashion [6]. When an active phishing warning is presented, a security decision is required. Because users' primary tasks are still online activities and security is only a secondary concern, users will put forth minimum effort to security. Under this circumstance, the handy skills can balance the users' effort and risk tradeoff, and are more likely to be applied for a security decision by the users [36].

Limitations. In our field experiment, the effectiveness of combining phishing warning and phishing training was evident with a sample drawn from university students, who tend to be younger, more educated, and more tech-savvy than the general public. A previous study [32] showed that younger participants are more susceptible to phishing than other age groups; thus the results obtained in the current study could be a low boundary if all the users are taken into consideration. Also, young people are faster in learning most tasks that require mental effort or demand attention [12]; thus the effectiveness of the phishing training on typical users might be over-estimated in the study. In addition, the phishing attack was simulated within an e-commerce context in our study. Due to user behavior differences found previously [30], different behavior might be observed if some other targeted account is chosen.

8. CONCLUSION

In this paper, we reported a field experiment, in which a simulated phishing attack was implemented to examine the effectiveness of an anti-phishing warning system. We chose to use traffic ranking as the criterion for phishing detection, and presented it as the reason why the warning was presented in the warning interface. The results showed that a combination of training regarding how to identify phishing URLs and our warning significantly reduced the number of users who were phished. However, training by itself was not sufficient: Participants who did not see warning after entering information all fell for the simulated phishing, regardless of whether training had been provided or not. Based on our results, we suggest integrating training in the warning interface. Also, we think involving traffic ranking in phishing detection, and explaining why warnings are generated will improve current phishing warning design.

Acknowledgement This paper is based upon work supported by the United States National Science Foundation under Grant No. 1314688 and by a National Security Agency Grant as part of a Science of Security lablet through North Carolina State University.

9. REFERENCES

[1] 2014a. Global Phishing Survey 1H2014: Trends and Domain Name Use. (2014). https://docs .apwg .org/ reports/ APWG _Global _Phishing _Report _1H _2014.pdf.

[2] 2014b. Phishing Activity Trends Report. (2014). https://docs .apwg .org/ reports/ apwg _trends _report _q2 _2014.pdf.

[3] 2015. Email authentication. (2015). https://support. google. com/ mail /answer /180707?hl=en.

[4] 2016. Chrome Privacy White Paper. (2016). https://www.google.com /chrome /browser /privacy /whitepaper.html.

[5] Devdatta Akhawe and Adrienne Porter Felt. 2013. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness.. In *Usenix Security*. 257–272.

[6] John R Anderson. 1983. *The architecture of cognition*. Psychology Press.

[7] Lujo Bauer, Cristian Bravo-Lillo, LF Cranor, and Elli Fragkaki. 2013. Warning design guidelines. *Pittsburgh, PA: Carnegie Mellon University* (2013).

[8] L. E. Bourne and A. F. Healy. 2012. Training and Its Cognitive Underpinnings. In *Training cognition: Optimizing Efficiency, Durability, and Generalizability*, A. F. Healy and L. E. Bourne (Eds.). Psychology Press.

[9] Cristian Bravo-Lillo, Lorrie Cranor, Julie Downs, Saranga Komanduri, Stuart Schechter, and Manya Sleeper. 2012. Operating system framed in case of mistaken identity: measuring the success of web-based spoofing attacks on OS password-entry dialogs. In *Proceedings of the 2012 ACM conference on computer and communications security*. ACM, 365–377.

[10] Neil Chou, Robert Ledesma, Yuka Teraguchi, John C Mitchell, and others. 2004. Client-Side Defense Against Web-Based Identity Theft.. In *NDSS*.

[11] Jason W Clark and Damon McCoy. 2013. There Are No Free iPads: An Analysis of Survey Scams as a Business.. In *LEET*.

[12] Fergus IM Craik and Janine M Jennings. 1992. *Human memory*. Lawrence Erlbaum Associates, Inc, Chapter Handbook of aging and cognition.

[13] Lorrie Faith Cranor. 2008. A Framework for Reasoning About the Human in the Loop. *UPSEC* 8 (2008), 1–15.

[14] Rachna Dhamija and J Doug Tygar. 2005. The battle against phishing: Dynamic security skins. In

Proceedings of the 2005 symposium on Usable privacy and security. ACM, 77–88.

[15] Julie S Downs, Barbagallo Donato, and Acquisti Alessandro. 2015. Predictors of risky decisions: Improving judgment and decision making based on evidence from phishing attacks. In *Neuroeconomics, judgment, and decision making*, Evan A Wilhelms and Valerie F Reyna (Eds.). Psychology Press, 239–253.

[16] Julie S Downs, Mandy B Holbrook, and Lorrie Faith Cranor. 2006. Decision strategies and susceptibility to phishing. In *Proceedings of the second symposium on Usable privacy and security*. ACM, 79–90.

[17] W Keith Edwards, Erika Shehan Poole, and Jennifer Stoll. 2008. Security automation considered harmful?. In *Proceedings of the 2007 Workshop on New Security Paradigms*. ACM, 33–42.

[18] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 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*. ACM, 1065–1074.

[19] Adrienne Porter Felt, Alex Ainslie, Robert W Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. 2015. Improving SSL Warnings: Comprehension and Adherence. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2893–2902.

[20] J Paul Frantz. 1994. Effect of location and procedural explicitness on user processing of and compliance with product warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 36, 3 (1994), 532–546.

[21] Stefan Görling. 2006. The myth of user education. In *Virus Bulletin Conference*, Vol. 11. 13–16.

[22] Amir Herzberg and Ahmad Gbara. 2004. *Trustbar: Protecting (even naive) web users from spoofing and phishing attacks*. Technical Report. ryptology ePrint Archive, Report 2004/155. http://eprint. iacr. org/2004/155.

[23] Robert R Hoffman. 2014. *The psychology of expertise: Cognitive research and empirical AI*. Psychology Press.

[24] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. 2009. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security*. ACM, 3–15.

[25] Kenneth R Laughery and Michael S Wogalter. 2006. Designing effective warnings. *Reviews of human factors and ergonomics* 2, 1 (2006), 241–271.

[26] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. 2011. Does domain highlighting help people identify phishing sites?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2075–2084.

[27] Christian Ludl, Sean McAllister, Engin Kirda, and Christopher Kruegel. 2007. On the effectiveness of techniques to detect phishing sites. In *DIMVA*, Vol. 7. Springer, 20–39.

[28] Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. 2009a. Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1245–1254.

[29] Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. 2009b. Identifying suspicious URLs: an application of large-scale online learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 681–688.

[30] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, and Cate Jerram. 2015. The design of phishing studies: Challenges for researchers. *Computers & Security* (2015).

[31] Robert W Proctor and Addie Dutta. 1995. *Skill acquisition and human performance*. Sage Publications, Inc.

[32] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. 2010. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 373–382.

[33] Steve Sheng, Brad Wardman, Gary Warner, Lorrie Cranor, Jason Hong, and Chengshan Zhang. 2009. An empirical analysis of phishing blacklists. In *Sixth Conference on Email and Anti-Spam (CEAS)*. California, USA.

[34] Andreas Sotirakopoulos, Kirstie Hawkey, and Konstantin Beznosov. 2011. On the challenges in usable security lab studies: lessons learned from replicating a study on SSL warnings. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*. 3–15.

[35] Colin Whittaker, Brian Ryner, and Marria Nazif. 2010. Large-Scale Automatic Classification of Phishing Pages.. In *NDSS*, Vol. 10.

[36] Christopher D Wickens. 2014. Effort in human factors performance and decision making. *Human Factors: The Journal of the Human Factors and Ergonomics Society* (2014), 1–8.

[37] Michael S Wogalter, Dave DeJoy, and Kenneth R Laughery. 2005. *Warnings and risk communication*. CRC Press.

[38] Michael S Wogalter, Russell J Sojourner, and John W Brelsford. 1997. Comprehension and retention of safety pictorials. *Ergonomics* 40, 5 (1997), 531–542.

[39] Min Wu, Robert C Miller, and Simson L Garfinkel. 2006. Do security toolbars actually prevent phishing attacks?. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 601–610.

[40] Guang Xiang, Jason Hong, Carolyn P Rose, and Lorrie Cranor. 2011. Cantina+: A feature-rich machine learning framework for detecting phishing web sites. *ACM Transactions on Information and System Security (TISSEC)* 14, 2 (2011), 21.

[41] Yue Zhang, Serge Egelman, Lorrie Cranor, and Jason Hong. 2006. Phinding phish: Evaluating anti-phishing tools. ISOC.