The Importance of Being Earnest [in Security Warnings]

Serge Egelman[a] and Stuart Schechter[b]

[a]University of California, Berkeley
[b]Microsoft Research, Redmond
egelman@cs.berkeley.edu
stuart.schechter@microsoft.com

Abstract. In response to the threat of phishing, web browsers display warnings when users arrive at suspected phishing websites. Previous research has offered guidance to improve these warnings. We performed a laboratory study to investigate how the choice of background color in the warning and the text describing the recommended course of action impact a user's decision to comply with the warning. We did not reveal to participants that the subject of the study was the warning, and then we observed as they responded to a simulated phishing attack. We found that both the text and background color had a significant effect on the amount of time participants spent viewing a warning, however, we observed no significant differences with regard to their decisions to ultimately obey that warning. Despite this null result, our exit survey data suggest that misunderstandings about the threat model led participants to believe that the warnings did not apply to them. Acting out of bounded rationality, participants made conscientious decisions to ignore the warnings. We conclude that when warnings do not correctly align users' risk perceptions, users may unwittingly take avoidable risks.

1 Introduction and Background

Many web browsers use full screen warning messages that are displayed to users whenever they visit suspected phishing websites. Egelman et al. studied several of these warnings and proposed a set of recommendations for improving them [1]. These recommendations included designing warnings that get noticed by interrupting the user's primary task, recommending a clear course of action so that the user knows what to do, distinguishing them from less serious warnings to prevent habituation, and minimizing the impact that a well-designed forgery has on a user's trust. In this study, we performed a controlled experiment to examine some of these recommendations.

The first question we examined was whether clearer explanations of users' available options would result in them making better choices. Most browser-based phishing warnings present users with multiple options, usually a recommendation not to visit a suspicious website and another option to bypass the warning. We examined the options offered by Microsoft's Internet Explorer 8 phishing warning [3].

When a user visits a suspected phishing website, she is advised to "go to my homepage instead." Because this text does not conceptually help the user complete her primary task—it was unlikely that she was trying to visit her homepage—we were concerned that this text may contribute to the warning being ignored. Thus, we tested how option text impacts decisions by creating an experimental condition that appeared to be more likely to aid in completing the primary task: "search for the real website." We hypothesized that this text would be more effective because it may facilitate completion of a primary task and it underscores the threat model: the user was visiting a fraudulent website designed to look like a legitimate one and therefore following this link may help the user locate the intended website. In addition to examining the option text, we wanted to examine the recommendation to minimize habituation by designing phishing warnings differently from less-severe warnings. Thus, we also varied the background color by turning it red in some conditions, while keeping it white in another.

We contribute to the literature on security warnings by showing that altering text and color significantly increase user attention. However, we show that attention alone is insufficient for warning compliance; because many participants did not believe the warnings were relevant to them, they chose to ignore them. We conclude that a user may face moral hazard when she encounters a security warning that does not effectively communicate the risk it is trying to mitigate. We later validated this finding in subsequent work [4].

2 Methodology

We performed our experiment on the Microsoft campus, using a recruitment service to obtain a (non-university-biased) sample of 59 participants. We did not tell participants that we were studying security, as that would compromise external validity by priming them. Instead, we told them that we were performing a usability study of Hotmail and therefore only recruited Hotmail users. At the time, Hotmail was the largest webmail provider worldwide [5], and therefore we believe our sample is generalizable to a large proportion of Internet users.

We randomly assigned participants to one of three between-group conditions and then gave them a set of tasks that involved checking email. After completing the final task, participants received a simulated phishing email. We observed their reactions to a warning from one of the three treatments and then asked them to fill out an exit survey.

2.1 Recruitment

We recruited participants during September of 2008. Thirty were male and the mean age was 37.6 ($\sigma = 11.6$). We selected participants who had previously opted in to being contacted about user studies at Microsoft, and screened out participants who either did not use Hotmail for their email or IE as their primary web browser. Because we were only interested in participants who were most vulnerable to phishing attacks, we screened out participants who had technical jobs.

When a participant arrived for his individual session, we asked him to sign a consent form, and then handed him an instruction sheet. The experimenter read the instructions aloud to ensure that everyone received the same information. When ready to begin, the experimenter left the room so as to not influence participants' behaviors. The experimenter observed participants from a separate control room as they completed a series of tasks. Once complete, the experimenter returned to administer an exit survey.

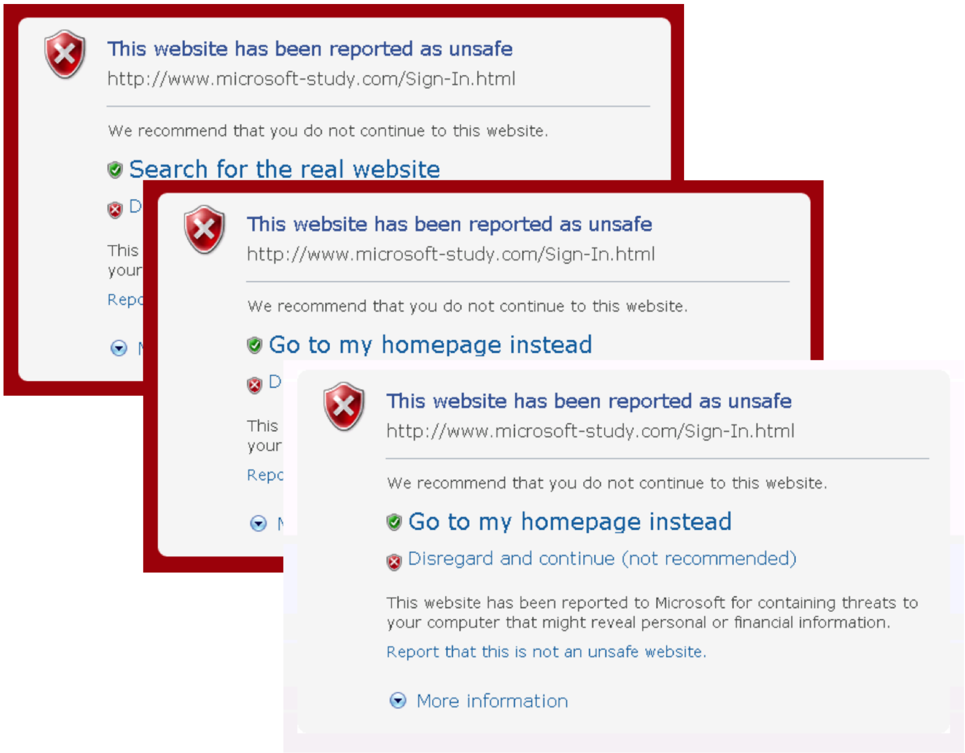
2.2 Tasks

To maximize ecological validity, we wanted participants to behave as they would when seeing phishing warnings outside the laboratory. Since security is rarely a primary task (e.g., users do not sit down at the computer to "not get phished"), we needed to mask the purpose of the study. We told participants that we were examining the usability of Hotmail. As an incentive, we paid them a dollar for each message that they opened during the study, and an additional four dollars if they "interacted with that message." So as to not bias them towards our phishing messages—we did not want them to feel compelled to click links in every message—we told them that filing messages away or deleting them would also count. Thus, we created an incentive only to read every received email, not necessarily to follow its instructions; participants received just as much compensation for deleting a message as following its links.

Because we could not solely rely on them to receive real email during the study period from outside sources, we explained that the experimenter would send them a message every ten minutes, but did not specify how many times this would occur. The first message sent by the experimenter was a personal message written in plaintext that asked the participant to visit a movie website and respond with the movie they most wanted to see. The second message was an HTML-based message that came from a photo-sharing website inviting the participant to view a shared photo album that the experimenter had posted. These two messages served only to further convince participants that they were part of an email usability study, we therefore do not mention them again.

Two minutes after participants viewed the second email message, the experimenter sent a simulated phishing message. This message did not follow the ten minute interval and was intended to create some ambiguity as to whether it was part of the study or not. The domain that we used to send it was registered solely for the study, though in its body it claimed to be from Microsoft and encouraged readers to click a link and enter personal information on the resulting website. The domain used for the destination URL as well as sending the email, *microsoft-study.com*, was added to a phishing blacklist, thereby triggering a phishing warning when accessed. The message offered participants the opportunity to enter a prize drawing if they visited the included URL. Upon arriving at this URL, participants saw one of three warnings that we describe in the next section. If they chose to ignore the warning and proceed to the website, they were presented with a login form that appeared identical to the Hotmail login screen (i.e., the goal of the simulated phishing scam was to capture Hotmail credentials).

In real life, a phishing warning appears after a user has clicked a link in a scam email. For ecological validity, we needed participants to be in this same frame of mind, which is why we incentivized them to read messages received during the study. Specifically, if a participant did not read the message, she would never attempt to visit the suspicious website, she would never see one of the three warning messages, and we would not yield any data regarding whether or not she would have obeyed the warning. We were not measuring how many messages participants read or how many websites they visited. The behavior we were studying was whether, after reading a scam email and visiting its included URL, participants would dismiss the phishing warning and submit login credentials. Thus, our dependent variable was whether participants entered information into the fake Hotmail login website.



This website has been reported as unsafe
http://www.microsoft-study.com/Sign-In.html

We recommend that you do not continue to this website.

Search for the real website

D

This
your

Repo

M

This website has been reported as unsafe
http://www.microsoft-study.com/Sign-In.html

We recommend that you do not continue to this website.

Go to my homepage instead

D

This
your

Repo

M

This website has been reported as unsafe
http://www.microsoft-study.com/Sign-In.html

We recommend that you do not continue to this website.

Go to my homepage instead

Disregard and continue (not recommended)

This website has been reported to Microsoft for containing threats to your computer that might reveal personal or financial information.

Report that this is not an unsafe website.

More information

Fig. 1. Participants who clicked the link contained in the simulated phishing email were exposed to one of three possible phishing warnings. The bottom represents the *search* condition, while the middle represents the *home* condition, and the top represents the *white* condition.

2.3 Conditions

We created our initial two conditions to examine the role of the option text: one warning recommended that users "go to my homepage instead," while the warning in the other condition recommended that they "search for the real website." We refer to these as the *home* and *search* conditions, respectively. Our hypothesis was that study participants would be less likely to heed the recommendations of the phishing warnings if those recommendations appeared unlikely to help complete a primary task (i.e., they were not attempting to visit a homepage at the time that the warning appeared). We believed that the text "search for the real website" would not conflict with the primary task as well as underscore the threat model.

Previously, Egelman et al. concluded that users were habituated to ignoring the IE7 phishing warnings because these warnings were designed similarly to other IE7 security warnings, such as those used to indicate SSL errors [1]. We created a third condition to examine habituation effects by removing the red border, and replacing it with a white border, so that it would look similar to the ubiquitous IE7 warnings. We refer to this condition as the *white* condition. Our three experimental conditions are described in Table 1 and depicted in Figure 1. We intentionally did not create a fourth condition, to separate the effects of the red border from the effects of the new text (e.g., a warning with a white border using the "search for the real website" text), over concerns about statistical power. Specifically, we designed this experiment as a first inquiry into whether an effect exists, rather than an attempt to quantify the size of that effect.

Condition	Option Text	Background	Total Time	Average Views	Average Time
White	*Go to my homepage instead*	White	12.00s	1.36	9.76s
Home	*Go to my homepage instead*	Red	17.81s	1.67	10.76s
Search	*Search for the real website*	Red	30.97s	2.67	11.84s

Table 1. Descriptions of the three conditions as well as summary statistics for the total viewing time, average number of views per user, and the average time per view. Participants in the *search* condition viewed the warnings significantly more frequently as well as for significantly longer periods of time in total.

3 Results

We observed 48 of 59 participants (81%) follow the link to the suspected phishing website. Due to technical difficulties, three of these participants saw no phishing warnings and therefore proceeded to enter their personal information (i.e., in the absence of a warning, participants believed this was a legitimate website). Throughout the rest of this paper, we focus on the 45 remaining participants who saw one of the three phishing warnings. Of these 45 participants who viewed the warnings, twelve entered personal information (27%), whereas everyone else navigated away from the website.

A chi-square test did not show that participants in any one condition were significantly more likely to divulge their credentials: five in the *white* condition (33% of 15), three in the *home* condition (20% of 15), and four in the *search* condition (26.7% of 15). We believe that this null result has more to do with low statistical power stemming from our limited sample size. However, we found a significant interaction effect based on both the red border and the text of the warnings with regard to the amount of time participants spent reading the warnings. Table 1 lists the total time participants in each condition spent viewing the phishing warnings, the number of times they revisited the phishing warnings, and the average time spent viewing the warnings.[1]

We performed a Kruskall-Wallis one-way analysis of variance and found that participants in the *search* condition viewed the phishing warnings for significantly longer time in total ($\chi^2_2 = 7.83$, $p < 0.020$). Upon performing post-hoc analysis using a Mann-Whitney U test, we found that this was due to significant differences between the 31s average viewing time in the *search* condition and the 12s average viewing time

[1] We removed data from one participant in the *white* group after he—against directions—asked for help and then waited for the experimenter to respond from the observation room, therefore artificially increasing the amount of time he spent viewing the warning.

in the *white* condition ($p < 0.010$; Cohen's $d = 0.98$). Likewise, when examining the total number of times that participants viewed the warnings, we found that those in the *search* condition went back to review the warning significantly more often (i.e., they closed the warning, reread the email message, clicked the link again, etc.; $\chi^2_2 = 7.02$, $p < 0.030$). This was also attributed to the contrast with the *white* condition ($p < 0.012$; Cohen's $d = 0.99$). This indicates an interaction effect between the red background and the new text; participants spent significantly longer analyzing the warnings only when both these features were present.

Using our exit survey, we found a significant correlation between participants ignoring warnings during the experiment and claiming to have seen them previously ($\phi = 0.497$, $p < 0.001$); nine of the twelve "victims" said they recognized the warnings (75%), whereas only seven of the thirty-three non-victims (21%) claimed to have recognized the warnings. Thus, the combination of the new text and red background decreased habituation, which may explain why participants in the *search* condition spent significantly longer viewing the warnings.

4 Discussion

Our warning manipulations increased the amount of time participants spent reading the warnings. It is not clear whether the originality of the designs simply decreased habituation, or whether the new option text caused them to think more about their choices.[2] Still, a third of our participants ultimately succumbed to the attack. We found no correlation between falling for the attack and the amount of time spent viewing the warnings. Thus, while participants in the *search* condition paid more attention to the warnings, they were just as likely to dismiss them. In this section we discuss some possible reasons for why the warnings failed and how warning effectiveness may be improved.

4.1 Bounded Rationality

Of the twelve participants who divulged credentials, all but one understood that the warnings wanted them to navigate away (i.e., "do not visit the website"). The one participant, who was in the *white* condition, responded *"check the sender or link to make sure it would not be harmful."* Thus, participants did not disregard the warnings because they did not understand what the warnings wanted them to do. Instead, we believe that participants chose to disregard the warnings because they did not believe they were at risk; none of the warnings (Figure 1) mentioned a specific threat unless the user clicked the "more information" link. The warnings only said that the website "has been reported as unsafe" and that it was reported for "containing threats to your computer." These terms are vague and do not describe one specific threat model. Thus, it is not surprising that participants who ignored the warnings did not understand the threat: ten of the participants who ignored the warnings (83% of 12) said that they did so because they were visiting a legitimate website.

[2] Six (40% of 15) participants in the *search* condition attempted to use the search functionality of the warning to find the "real" website. Since no real website existed, this proved futile.

Users are exposed to many varying ill-defined online threats. In research on users' mental models of computer security, Wash observed that this has resulted in widely varying conceptualizations when given vague terms like "security" and "hacker" [6]. Because the warnings used terminology like "unsafe" and "threats to your computer," without providing details, participants likely had varying mental models. When given explanations of the threat model, participants acted rationally: nine of ten participants who clicked the "more information" link, and read about phishing, complied with the warning. These participants correctly understood that the website was attempting to steal their credentials.

4.2 Moral Hazard

We examined participants' understandings of the threat model by asking them to explain the danger of ignoring the warnings. We coded correct answers as ones that said phishing scams attempt to steal personal information. We found that only 14 understood this (31% of 45). Of the remaining 31 participants, all of them mentioned other threats. Some examples included:

- *"I could potentially get a virus or spyware"*
- *"Getting a virus ruining your computer"*
- *"Will get some spyware"*

Three participants who disregarded the warnings (25% of 12) said that they did not care if our computer was infected with a virus. That is, because they believed that someone else would bear all the risk from an infected computer, they did not believe there were any incentives to obeying the warnings. While this would be a rational justification if the threat were indeed malware (see, e.g., [2]), it illustrates how bounded rationality, caused by a limited understanding of the threat model, resulted in moral hazard.

4.3 Lack of Motivation

In Wogalter's Communication-Human Information Processing Model (C-HIP) [7], people undergo several steps between warning exposure and choosing an action. Motivation is a key step: users are unwilling to comply with warnings that they do not believe apply to them. Thus, the changes we made to the warning resulted in improvements at the attention stages of the model by minimizing habituation effects (this was corroborated by the significant correlation between participants ignoring the warnings and claiming to recognize them; those in the *search* condition were least likely to recognize them). However, the warnings failed because they failed to motivate participants.

We therefore believe that our experimental results indicate that motivation problems may be preventable by designing warnings to explicitly state a threat model. In fact, we later performed a followup experiment to validate this finding [4]: participants were significantly more likely to obey SSL warnings when those warnings explicitly communicated threat models that participants found to be relevant to them.

5 Conclusion

We expected to find that by using techniques to increase attention, participants would be more likely to obey the warnings because they would spend more time reading them. We found that we were partially correct: participants spent more time reading the warnings, but they ultimately did not behave any differently. Our exit survey data suggests participants who were unmotivated by the threat model—as they understood it—chose to disobey the warnings. We expected to observe a much greater effect size and therefore used a limited sample. In the *white* condition, ten of fifteen participants complied with the warnings. We consider this to be the baseline rate of compliance because this condition was designed to appear similar to previous phishing warnings (i.e., this condition approximated a control). Given our sample size, for there to be a significant difference between one of the other experimental conditions and this baseline rate of compliance (67%), the other conditions would need compliance rates of 100%. Thus, phishing warnings have improved to the point that much larger sample sizes are needed to quantitatively study minor design changes.

While we were unable to reject the null hypothesis, this study yielded important lessons for future security mitigations. We showed that distinguishing severe risks from other less-severe risks may aid in capturing user attention. However, warnings cannot rely on attention alone, they must also communicate risk effectively. Many participants incorrectly believed they were being warned about different irrelevant threats. In future warnings, designers should highlight the risks of ignoring the warnings so that users are more likely to understand that the warnings apply to them. This means warning less often in low risk situations, providing stronger evidence of the presence of risk, or helping users to link the risk to their immediate situations through contextual cues.

References

1. S. Egelman, L. F. Cranor, and J. Hong. You've been warned: An empirical study of the effectiveness of web browser phishing warnings. In *CHI '08: Proceeding of The 26th SIGCHI Conference on Human Factors in Computing Systems*, pages 1065–1074, New York, NY, USA, 2008. ACM.
2. C. Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *New Security Paradigms Workshop*, pages 133–144, 2009.
3. E. Lawrence. IE8 Security Part III: SmartScreen Filter, July 2008. http://blogs.msdn.com/ie/archive/2008/07/02/ie8-security-part-iii-smartscreen-filter.aspx.
4. J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying wolf: an empirical study of ssl warning effectiveness. In *Proceedings of the 18th USENIX Security Symposium*, SSYM'09, pages 399–416, Berkeley, CA, USA, 2009. USENIX Association.
5. D. Terdiman. Microsoft aiming to clean up hotmail user's inboxes. CNET News, October 3 2011. `http://news.cnet.com/8301-13772_3-20114975-52/microsoft-aiming-to-clean-up-hotmail-users-inboxes/`.
6. R. Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, New York, NY, USA, 2010. ACM.
7. M. S. Wogalter. Communication-Human Information Processing (C-HIP) Model. In M. S. Wogalter, editor, *Handbook of Warnings*, pages 51–61. Lawrence Erlbaum Associates, 2006.