User Perceptions of Phishing Consequence Severity and Likelihood, and Implications for Warning Message Design

Eleanor K. Foster[1(✉)], Keith S. Jones[1], Miriam E. Armstrong[1], and Akbar S. Namin[2]

[1] Department of Psychological Sciences, Texas Tech University, Lubbock, TX, USA
{eleanor.foster,keith.s.jones,miriam.armstrong}@ttu.edu
[2] Department of Computer Science, Texas Tech University, Lubbock, TX, USA
akbar.namin@ttu.edu

Abstract. To combat phishing, system messages warn users of suspected phishing attacks. However, users do not always comply with warning messages. One reason for non-compliance is that warning messages contradict how users think about phishing threats. To increase compliance, warning messages should align with user perceptions of phishing threat risks. How users think about phishing threats is not yet known. To identify how users perceive phishing threats, participants were surveyed about their perceptions of the severity and likelihood of 9 phishing consequences. Results revealed perceived severity and likelihood levels for each consequence, as well as relative differences between consequences. Concrete examples of warning messages that reflect these findings are provided.

Keywords: Human factors · Cybersecurity · Social engineering · Phishing · Consequence · Severity · Likelihood · Risk · Warning · Message

1 Introduction

Phishing occurs when someone attempts to obtain sensitive information through email. One strategy to help thwart phishing attacks is to warn users when an attack is suspected. However, users do not always comply with warnings [e.g., 1]. Thus, researchers have studied how to design cybersecurity warnings to increase compliance.

That literature has produced three key recommendations. First, warnings should describe attack consequences [e.g., 2, 3]. For example, a message in response to an email asking the user to provide credit card information could warn that providing the requested information would enable the recipient to freely use their credit card to make any number of purchases and for any amount. Second, warnings should convey attack risk [e.g., 2, 4], which is a function of two factors: 1) the severity of the attack, and 2) the likelihood the user will experience the attack [2]. Continuing the previous example, the warning could convey that doing what the sender asked would be high risk because: 1) it would take a lot of time and effort to work with the credit card company to deal with fraudulent purchases (severity), and it is very likely the email is a phishing attempt

M. Zallio et al. (Eds.): AHFE 2021, LNNS 268, pp. 265–273, 2021.
https://doi.org/10.1007/978-3-030-79997-7_33

(likelihood). Third, warning messages should align with how users think about attacks [e.g., 5, 6]; otherwise, users will not trust the message [3, 7]. Continuing the previous example, the warning message could merely be descriptive if the user thinks attack risk is low, e.g., "it will be necessary to disavow any fraudulent purchases". Alternatively, the warning message could be more strongly worded if the user thinks attack risk is high, e.g., "it will be necessary to disavow however many fraudulent purchases the recipient makes, which could require a lot of time and effort".

To create a phishing warning message that accounts for all three recommendations, one would need to describe the risk associated with the phishing attack's consequences, and in a way that aligns with how users think about that risk. Research has investigated how users think about topics related to cybersecurity [e.g., 8], but not phishing attack consequence risk. Additional research is necessary if we are to design phishing warning messages that comply with all three of the design recommendations described above.

1.1 The Present Study

We investigated how users think about risks associated with phishing attack consequences. To do so, users rated the severity and likelihood of phishing attack consequences. We then analyzed those ratings to understand a) the level of perceived severity and likelihood for each consequence, and b) whether perceived severity, likelihood, or both varies across consequences. Finally, we offered concrete recommendations regarding how to apply our findings to the design of phishing warning messages.

2 Method

2.1 Survey

The survey consisted of a brief definition of phishing and two sets of questions. One set concerned the severity of 9 common phishing attack consequences (C1 through C9), each rated via a 7-point response item (1 = Not Severe, 7 = Severe); the other set concerned the likelihood of those consequences, each rated via a 7-point response item (1 = Not at all likely, 7 = Very likely). Table 1 provides descriptions for each consequence.

C1 and C2 concern situations in which users do what the phisher asked, but are not aware of consequences other than perhaps being tricked. Questions about C1 and C2 allowed us to assess how users perceive the risk of clicking a phishing link or providing personal information per se. C3 through C9 concern situations in which users experience consequences beyond being tricked. Questions about C3 through C9 allowed us to assess how users perceive the risk associated with each of those consequences.

2.2 Procedure

Each participant 1) provided informed consent, 2) completed demographic questions, 3) completed the survey, which was embedded within a larger survey and administered online, and 4) received partial course credit. The research complied with the APA Code of Ethics, and was approved by the Texas Tech Institutional Review Board.

Table 1. The 9 common phishing attack consequences investigated.

Consequence	Description
C1	Phished because clicked phishing link
C2	Phished because gave phisher personal information via email
C3	Phisher gains your username & password for Web site
C4	Phisher performs actions on a Web site as if they were you
C5	Phisher accesses your information stored in a Web site
C6	Phisher deletes your information stored in a Web site
C7	Phisher modifies your information stored in a Web site
C8	Phisher prevents you from logging into a Web site
C9	Phisher takes control over one of your financial accounts

2.3 Participants

In total, 1,649 students in an Introduction to Psychology course completed the study. Their data were examined for missing responses, careless responding, and outlier response patterns. Cases missing responses to more than 10% of our survey items were identified and removed. One hundred and twelve participants were removed for missing data. We then employed a long strings evaluation [9] to identify and remove participants from the data set who rated all response items the same. Five hundred two participants were removed for careless responding. Last, to identify participants whose response patterns were particularly unusual (chi-square $p < .001$), we computed Mahalanobis Distance [11] for the severity and susceptibility response items separately. One hundred twenty-one cases were removed from the data set for being outliers. After completing these steps, 914 participants (659 females, 252 males, 3 other) remained in the data set. Their ages ranged from 16–49 years ($M = 19.03$, $SD = 2.52$).

3 Results

3.1 Replacing Small Amounts of Missing Data

We retained cases missing fewer than 10% of response items. For those cases, we employed 2 methods to replace the missing data. First, we used hot decking [10] to replace missing data with values from another participant whose responses were identical to the participant's non-missing data (via the SPSS macro). Second, if no donor case was found, we replaced missing data with the mean of the missing data point. Twenty-one data points, .001% of the dataset, were replaced using these methods.

3.2 What Were the Perceived Severity and Likelihood Levels?

We used bootstrapping [12] (1000 samples; sampled with replacement; sample size = 914) to compute a mean and confidence interval for perceived severity and likelihood for

each consequence. We did so, rather than computing those statistics for the sample as a whole, to provide the best possible estimate of the population mean for each consequence. Table 2 provides the resultant means and confidence intervals.

Inspection of Table 2 suggests perceived severity ranged from moderate to severe, with all consequences except C1 (Phished because clicked phishing link) falling in the fairly severe (above the mid-point but below the top-end) to severe range. In contrast, perceived likelihood fell in the somewhat likely range (above the bottom-end but below the mid-point); no consequences were rated as moderately likely to occur or greater.

Table 2. Perceived severity and likelihood rating means and confidence interval estimates

Consequence	Mean Severity	95% CI Severity	Mean Likelihood	95% CI Likelihood
C1	3.87	[3.80, 4.00]	3.18	[3.08, 3.32]
C2	5.40	[5.29, 5.51]	2.61	[2.47, 2.73]
C3	5.69	[5.60, 5.80]	3.18	[3.08, 3.32]
C4	6.16	[6.11, 6.29]	3.15	[3.07, 3.33]
C5	6.01	[5.91, 6.09]	3.24	[2.97, 3.23]
C6	5.47	[5.40, 5.60]	3.01	[3.08, 3.32]
C7	5.92	[5.81, 5.99]	3.12	[2.88, 3.12]
C8	5.66	[5.60, 5.80]	3.24	[2.98, 3.22]
C9	6.68	[6.63, 6.77]	3.18	[3.06, 3.34]

3.3 Do Perceived Severity or Likelihood Ratings Vary Across Consequences?

We examined differences in perceived risk between the nine consequences to investigate whether any consequences were perceived as more severe or more likely than others. We focused on differences because ratings were not independent [13].

To guard against Type I error, we randomly divided the data set into two sub-sets of 457 participants (Split 1 and Split 2) and performed this analysis on each sub-set. For each participant within each split, we then computed difference scores for each consequence pair (e.g., C1–C2), separately for severity and likelihood. To obtain estimates for each difference score pair, we ran a bootstrap with replacement using 1000 replications, sample sizes of 914, and a corrected alpha of .001. We employed a stringent 99.9% confidence interval because we considered the 36 differences associated with each rating type (severity or likelihood) to be a family and aimed to maintain family-wise error at .05 (alpha = .05/36 = .001). In the following paragraphs, we interpret only effects observed in both splits.

3.3.1 Perceived Severity for Pairs of Consequences

Table 3 provides confidence intervals for the 30 severity difference scores that were statistically significant in Split 1 *and* 2. Table 3 reveals 1) C9 (Phisher takes control

over one of your financial accounts) was rated as significantly *more* severe than all other consequences, 2) C4 (Phisher performs actions on a Web site as if they were you), C7 (Phisher modifies your information stored in a Web site), and C5 (Phisher accesses your information stored in a Web site) were rated as significantly *more* severe than C2 (Phished because gave phisher personal information via email), and 3) all consequences were rated as more severe than C1(Phished because clicked phishing link) (Table 4).

4 Discussion

We had 2 goals: to determine 1) the levels of perceived severity and likelihood for each phishing consequence, and 2) whether individuals rated consequences as more or less severe, and more or less likely to occur. Our findings related to each will be described in the following sub-sections, followed by concrete examples of how our findings can be applied to warning message design.

4.1 Perceived Severity

Consequences fell into one of 3 groups: 1) severe, 2) fairly severe, and 3) moderately severe. Perceptions of severity appear to reflect the extent to which a consequence is contextualized and concrete.

The first group was comprised of a single consequence, i.e., C9 (Phisher takes control over one of your financial accounts). Mean perceived severity for C9 (6.68) approached the top-end of the severity scale (7). Further, perceived severity for C9 was significantly greater than that for all other consequences. To describe the risk associated with C9 in a way that aligns with how users think about C9, one should describe the severity of C9 with very strong language, and that language should be stronger than the language used to describe other consequences.

The second group was comprised of 6 consequences, i.e., C2 through C8. Mean perceived severity for this group ranged from 5.40–6.16, which means these consequences were perceived as fairly severe. Perceived severity for consequences at the high end of that range differed significantly from that for the consequence on the low end of that range; however, each of those consequences were also not significantly different from other consequences in this group. That suggests perceived severity for consequences in this group were more homogeneous than not. To describe the risk associated with C2 through C8 in a way that aligns with how users think about those consequences, one should describe severity with fairly strong language, but that language should not be as strong as the language used to describe C9.

The third group was comprised of a single consequence, i.e., C1 (Phished because clicked phishing link). Mean perceived severity for C1 (3.87) approached the severity scale's mid-point (4); thus, C1 was perceived as moderately severe. Further, perceived severity for C1 was significantly less than that for all other consequences. This presents a challenge for describing the risk associated with C1. Security personnel do not want users to click links in phishing emails. Hoping to discourage users from doing so, they may create strongly worded messages warning users that something extremely bad could happen if they click a link. However, our results suggest such warnings will likely

Table 3. Significant 99.9% confidence intervals for severity difference scores.

Consequence pair	Split 1 (n = 457)	Split 2 (n = 457)
C1-C2	[-1.76, − 1.24]	[−1.75, − 1.25]
C1–C3	[−2.09, − 1.51]	[−2.07, − 1.53]
C1–C4	[−2.58, − 2.02]	[−2.58, − 2.02]
C1–C5	[−2.38, − 1.82]	[−2.49, − 1.91]
C1–C6	[−1.92, − 1.28]	[1.92, − 1.28]
C1–C7	[−2.29, − 1.71]	[−2.40, − 1.80]
C1–C8	[−2.11, − 1.49]	[−2.02, − 1.38]
C1–C9	[−3.09, − 2.51]	[−3.07, − 2.53]
C2–C3	[−.51, − .09]	[−.51, − .09]
C2–C4	[−.90, − .50]	[−1.01, − .59]
C2–C5	[−.83, − .37]	[−.81, − .39]
C2–C7	[−.74, − .26]	[−.83, − .37]
C2–C9	[−1.52, − 1.08]	[−1.53, − 1.07]
C3–C4	[−.56, − .24]	[−.65, − .35]
C3–C5	[−.46, − .14]	[−.46, − .14]
C3–C7	[−.38, − .02]	[−.47, − .13]
C3–C9	[−1.20, − .80]	[−1.19, − .81]
C4–C5	[.06,−.34]	[.06, .34]
C4–C6	[.50,−.90]	[.51, .89]
C4–C7	[.05, .35]	[.05, .35]
C4–C8	[.21, .59]	[.41, .79]
C4–C9	[−.65, − .35]	[−.63, − .37]
C5–C6	[.31, .69]	[.30, .70]
C5–C8	[.12, .48]	[.21, .59]
C5–C9	[−.87, − .53]	[−.75, − .45]
C6–C7	[−.65, − .35]	[−.64, − .36]
C6–C9	[−1.42, − .98]	[−1.41, − .99]
C7–C8	[.04, .36]	[.14, .46]
C7–C9	[−.97, − .63]	[−.86, − .54]
C8–C9	[−1.20, − .80]	[−1.30, − .90]

engender distrust because they will not align with how users think about clicking a potential phishing link [e.g., 3]. Alternatively, one could describe the severity of C1 with moderately strong language that is less strong than the language used to describe all other consequences, which would align with how users think about C1. That should

Table 4. Significant 99.9% confidence intervals for likelihood difference scores.

Consequence pair	Split 1 (n = 457)	Split 2 (n = 457)
C1–C2	[.25,.75]	[.35,.85]
C2–C3	[−.71, − .29]	[−.82, − .38]
C2–C4	[−.61, − .19]	[−.83, − .37]
C2–C5	[−.82, − .38]	[−.94, − .46]
C2–C6	[−.62, − .18]	[−.62, − .18]
C2–C7	[−.72, − .28]	[−.73, − .27]
C2–C8	[−.84, − .36]	[−.93, − .47]
C2–C9	[−.74, − .26]	[−.85, − .35]
C3–C6	[.04,.36]	[.04,.36]
C5–C6	[.06,.34]	[.15,.45]
C6–C8	[−.35, − .05]	[−.36, − .04]

increase the likelihood that users will trust the message [3, 5]. However, that increase in trust may not translate into compliance if the consequences of *not* clicking the link are perceived as more severe than the potential consequences of clicking it [5]. In such cases, it may be best to focus the wording of the warning message, not on clicking the link, but rather on what could happen if they do what the phisher asked. For example, one could word the warning message to convey that doing what the phisher asked could allow them to gain your username and password, which users perceive as fairly severe. Doing so would allow for the use of stronger language, which hopefully will convince users that the potential consequences of *not* doing what the phisher asked are less severe than the potential consequences of doing what they asked.

4.2 Perceived Likelihood

Certain consequences differed from one another, but all fell below the mid-point of the scale. Thus, all consequences were perceived as being only somewhat likely to occur. This presents another challenge for describing the risks associated with these consequences. Specifically, to describe risk in a way that aligns with how users think about those consequences, one should convey that these consequences are only somewhat likely to occur (regardless of the actual likelihood that the user will experience those consequences). Doing so should increase users' trust in the warning message [e.g., 3]. However, it will also probably decrease users' motivation to do what is required to prevent the attack [14]. Alternatively, one could ignore the recommendation to describe risk in a way that aligns with how users think about those consequences, and instead describe the actual likelihood that the user will experience those consequences [2, 5]. However, as noted earlier, that should decrease users' trust in the warning message [3] when attack likelihood is moderate to high. Accordingly, either of those approaches may do more harm than good. Therefore, it may be best to simply not describe likelihood in

warning messages [c.f., 2, 5]. That would avoid decreasing a) users' motivation to do what is required to prevent the attack or b) their trust in the warning message. As such, not describing consequence likelihood may be the lesser of the evils.

4.3 Concrete Examples of Warning Messages that Reflect Our Findings

Figure 1 provides concrete examples of messages that reflect our recommendations for warning users about C9 (Phisher takes control over one of your financial accounts), C3 (Phisher gains your username & password for Web site), and C1 (Phished because clicked phishing link).

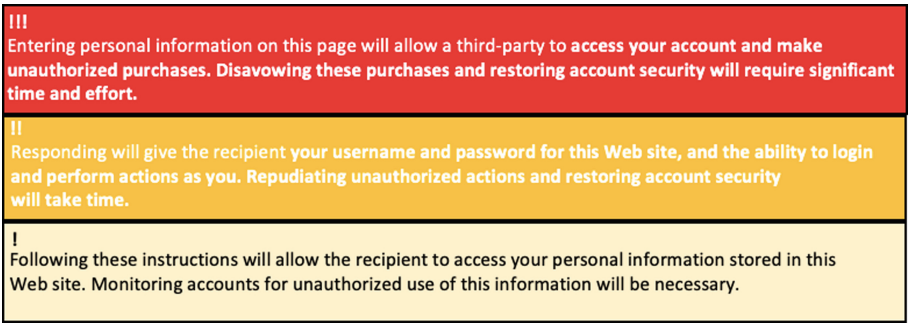

Fig. 1. Example warning messages reflecting our design recommendations.

Acknowledgements. This research was supported in part by the U.S. National Science Foundation (Award #: 1564293 & 1723765). Opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the NSF.

References

1. Bravo-Lillo, C., Cranor, L.F., Downs, J., Komanduri, S.: Bridging the gap in computer security warnings: a mental model approach. IEEE Sec. Priv. **9**, 18–26 (2011)
2. Hardee, J.B., West, R., Mayhorn, C.B.: To download or not to download: an examination of computer security decision making. Interactions **13**, 32–37 (2006)
3. Bartsch, S., Volkamer, M., Theuerling, H., Karayumak, F.: Contextualized web warnings, and how they cause distrust. In: Huth, M., Asokan, N., Čapkun, S., Flechais, I., Coles-Kemp, L. (eds.) Trust 2013. LNCS, vol. 7904, pp. 205–222. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38908-5_16
4. Bauer, L., Bravo-Lillo, C., Cranor, L.F., Fragkaki, E.: Warning design guidelines. CMU-CyLab. **13**, 1–27 (2013)
5. Bartsch, S., Volkamer, M.: Effectively communicate risks for diverse users: a mental-models approach for individualized security interventions. In: GI-Jahrestagung, pp. 1971–1984 (2013)
6. Blythe, J., Camp, L.J.: Implementing mental models. In: 2012 IEEE Symposium on Security and Privacy Workshops, pp. 86–90. IEEE Press, San Francisco (2012)

7. Ibrahim, T., Furnell, S.M., Papadaki, M., Clarke, N.L.: Assessing the usability of end-user security software. In: Katsikas, S., Lopez, J., Soriano, M. (eds.) TrustBus 2010. LNCS, vol. 6264, pp. 177–189. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15152-1_16
8. Wash, R.: Folk models of home computer security. In: Proceedings of the Sixth Symposium on Usable Privacy and Security, pp. 1–16. ACM, New York (2010)
9. Johnson, J.A.: Ascertaining the validity of individual protocols from web-based personality inventories. J. Res. Pers. **39**, 103–129 (2005)
10. Andridge, R.R., Little, R.J.A.: A review of hot deck imputation for survey non-response. Int. Stat. Rev. **78**, 40–64 (2010)
11. Meade, A.W., Craig, S.B.: Identifying careless responses in survey data. Psychol. Methods. **17**, 437–455 (2012)
12. Tabachnik, B.G., Fidell, L.S.: Using Multivariate Statistics, 6th edn. Pearson, Boston (2013)
13. Cumming, G., Finch, S.: Inference by eye: confidence intervals and how to read pictures of data. Am. Psychol. **60**, 170–180 (2005)
14. Krol, K., Moroz, M., Sasse, M.A.: Don't work. Can't work? Why it's time to rethink security warnings. In: 7th International Conference on CRiSIS, pp. 1–8. IEEE Press, Cork (2012)