A Follow-up Study of Detecting Phishing Emails

Ernst Bekkering, Ph.D.
bekkerin@nsuok.edu
Department of IS and Technology
Northeastern State University
Tahlequah, OK 74464, USA

Dan Hutchison
hutchisd@nsuok.edu
Department of IS and Technology
Northeastern State University
Tahlequah, OK 74464, USA

Laurie Werner
wernerla@muohio.edu
Department of Computer and Information Technology
Miami University Hamilton
Hamilton, OH 45011, USA

Abstract

Phishing continues to be an ongoing threat to online security. In a previous study, Werner and Courte (2008) demonstrated that training in detecting phishing emails helped students to feel that they were better able to deal with phishing attacks. This study follows up on that study and reports whether students were actually better at differentiating between phishing emails and legitimate emails after receiving training.

Keywords: phishing, information security, crimeware, computing literacy, prevention.

1. INTRODUCTION

Despite technological and educational countermeasures, phishing continues to be a continuing threat to online security. Recently, Consumer Reports estimated losses to phishing scams to be at almost a half-billion dollars over a two year period (2009a). Phishing and its more targeted version, spear phishing, continue to appear in the top 20 list of security threats (The SANS Institute, 2009). Spear phishing emails appear more credible because they contain information about specific staff or current organizational issues, which increases the appearance of legitimacy to members of the group or organization.

Though the number of reported phishing appears to be declining recently, the number of password-stealing URLs is on the rise (Anti-Phishing Working Group, 2009). Figure 1 illustrates the trends in reports to the Anti-Phishing Working Group (APWG) over the last four years.

Efforts to thwart phishing attacks generally fall into two categories: technological and educational. Internet browsers and email clients routinely contain specialized software to detect phishing. Internet Explorer 8 uses the Smartscreen filter, Firefox 3 enables built-in Phishing and Malware Protection by default, and Microsoft Outlook regularly updates its Junk E-mail filter. On the educational side, organizations like the FTC (Federal Trade Commission, 2006), the FDIC (Federal Deposit Insurance Corporation, 2008), Consumer Reports (Consumer Reports, 2009b), and multiple newspapers try to educate the public about detection of phishing emails. However, users frequently do not use the latest versions and updates of software. Analysis of recent browser use statistics demonstrates that of the top 10 browsers used, 68% is an older version and only 22% the latest browser version (W3counter, 2009) (**Table *1***). With regard to education, the situation is not much better. According to Jakobsson and Myers (2007), "a typical user does not know how to identify a phishing email". Participants in the 2008 Roundtable Discussion on Phishing Education agreed that "...more school-based education on computer security, cybersafety, and cyberethics is a good idea" (Federal Trade Commission, 2008).

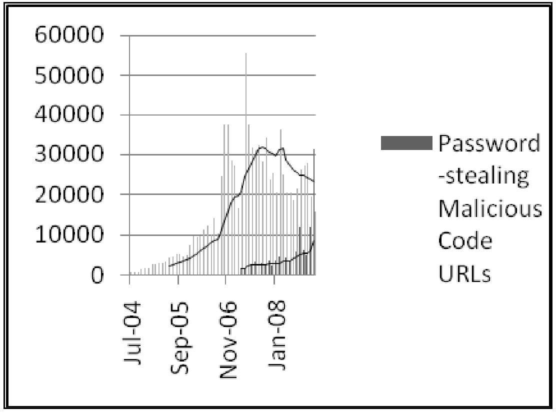
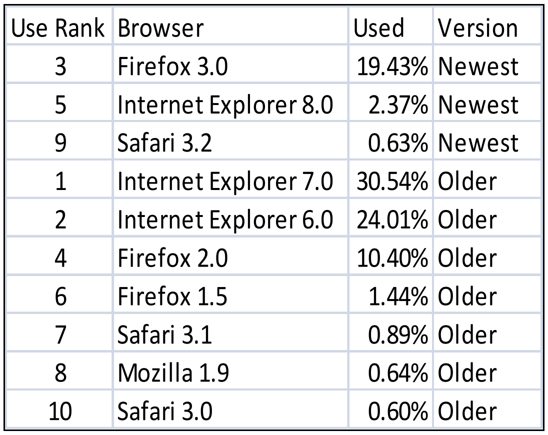


Figure 1- Trends Reported by Anti Phishing Working Group

In 2006, Robila and Ragucci (2006) conducted a study of effectiveness of education in teaching phishing detection. Participants in the study received instruction in the nature and detection of phishing emails, took a phishing IQ test customized by the authors, and completed a survey. On average, participants correctly identified 6.87 out of 12 emails. This barely exceeds pure chance, but students did report that they had become more aware of the issue of phishing. In 2007, Kamagura et al. (2007) used embedded training in a study with customized emails. Whenever a participant followed a link in a phishing email, the user would receive immediate feedback on phishing indicators. Using the immediate feedback, the authors demonstrated significantly improved efficacy and retention. Similar results were obtained in a follow-up study for spear phishing (Kumaraguru, Sheng, Acquisti, Cranor, & Hong, 2008). Following up on the Robila and Ragucci study in 2008, Werner and Courte (2008) reported the results of a lab activity using materials available to the general public. Students took a pre-test survey, then used the SonicWall Phishing IQ Test (Sonicwall Inc., 2008), and recorded their scores and perceptions in a post-test survey. The results indicated that using the SonicWall materials did make students feel more prepared to recognize phishing attacks, but did not measure actual performance since only a post-test was used, and students self-reported the results. This study is an extension of the Werner and Courte (2008) study, where performance is measured automatically before and after standardized instruction. The rest of the paper is structured as follows: we discuss the experimental design, then analyze the results, and close with our conclusions and recommendations.

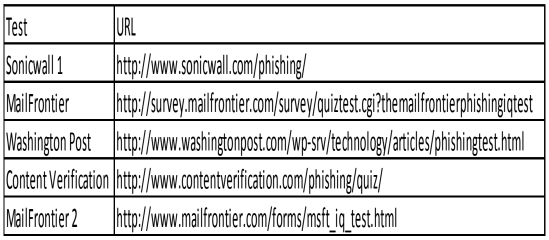


Use Rank	Browser	Used	Version
3	Firefox 3.0	19.43%	Newest
5	Internet Explorer 8.0	2.37%	Newest
9	Safari 3.2	0.63%	Newest
1	Internet Explorer 7.0	30.54%	Older
2	Internet Explorer 6.0	24.01%	Older
4	Firefox 2.0	10.40%	Older
6	Firefox 1.5	1.44%	Older
7	Safari 3.1	0.89%	Older
8	Mozilla 1.9	0.64%	Older
10	Safari 3.0	0.60%	Older

Table 1- Browser Versions Used

2. RESEARCH DESIGN

In this section, we first discuss how we structured our study to optimize reliability of the results. To measure performance in detecting of phishing emails, we generated two sets of ten emails. Each set contained five legitimate emails and five phishing emails to eliminate bias based on guessing. With a 50% score based on pure chance, participants would not have any benefit from an "all legitimate" or an "all phishing" strategy. Publicly available sources of "Phishing IQ tests" were used to select the twenty emails in the two sets. A listing of sources is provided in **Table *2***.



Test	URL
Sonicwall 1	http://www.sonicwall.com/phishing/
MailFrontier	http://survey.mailfrontier.com/survey/quiztest.cgi?themailfrontierphishingiqtest
Washington Post	http://www.washingtonpost.com/wp-srv/technology/articles/phishingtest.html
Content Verification	http://www.contentverification.com/phishing/quiz/
MailFrontier 2	http://www.mailfrontier.com/forms/msft_iq_test.html

Table 2- Sources of Phishing Tests

We further attempted to balance Phishing IQ test source and "companies" in order to form as identical tests as possible. For example, if three of the emails in set 1 came from MailFrontier, set 2 would contain three examples from MailFrontier too. Similarly, if set 1 contained and example using "Washington Mutual Bank", set 2 might contain an example based on "U.S. Bank". Half the participants used set 1 as a pretest, and the other half used set 2 as pretest. The sets were compared after data collection to check whether the results were similar enough to establish interchangeability. The opposite test would be used for the post test in both groups. For instance, if set 1 was used for the pretest, participants would receive set 2 as the post test and vice versa. Participants were randomly assigned to either combination of pretest and post test. Examples of legitimate and phishing emails *not* used in the two sets were used to generate the training materials. Consequently, there was no overlap between the pretest, post test, and training set at all. Pretest, training session, and post-test were posted on a secure website, and participants completed the three phases during class time with 2-3 days separation (**Figure 2**). Since participation was voluntary, and the participants were recruited from general IS courses rather than dedicated security courses, training and tests were not part of the regular curriculum. To enhance performance, extra credit points were offered for each email classified correctly in the pretest and the post test. Students only earned extra credit if they completed all three sessions, but an alternative activity was offered for those electing not to participate or unable to attend all three sessions. Finally, we explained to students that we did need their university ID to match the results on the pretest and the post test and to award the extra credit, but that we would not analyze their individual results.

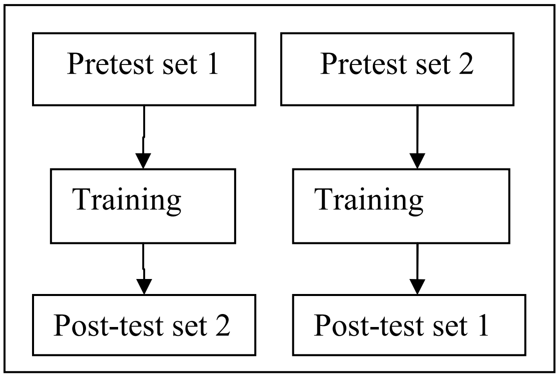
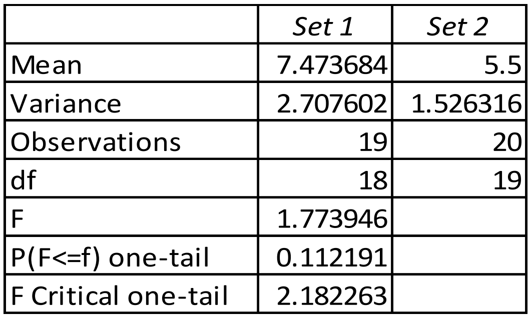


Figure 2 - Flow of Activities

3. DATA ANALYSIS

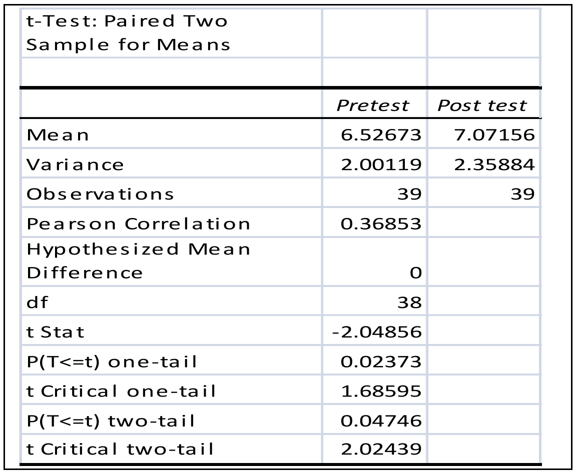
A total of 43 students participated in the study. Four students did not complete all three sessions, and were eliminated from the study. Distribution by gender was equal with 20 male and 19 female participants. The mean age was 23.7 years (s.d. = 7.0). Only three students listed "Information Systems" or "Computer Science" as their major. On the pretest and post-test, each correct answer was counted as one point for a total of 10 maximum. Eight pretests or post tests had a missing answer, and the scores on these tests were corrected by multiplying with 10/9.

Results of the data collection were analyzed using statistical tools in Microsoft Excel 2007, as well as the "Data Analysis Plus" add-on. First, we analyzed the 19 responses to set 1 and the 20 responses on set 2 pretests to test the assumption of equal variances. The result showed that the assumption of equal variances was met, but that the mean value showed significant differences (Table 3). Set 1 had a mean score of 7.5, whereas set 2 had a mean of 5.5. Since set 2 was clearly more difficult than set 1, scores for set 2 were increased with half the difference (+1) and scores for set 1 were decreased with the same amount (-1). After these corrections for non-identical tests, we ran a paired t-test to measure the effect of the instruction (Table 4). The results demonstrated a modest improvement in score from 6.5 to 7.1, but the result was statistically significant.



	Set 1	*Set 2*
Mean	7.473684	5.5
Variance	2.707602	1.526316
Observations	19	20
df	18	19
F	1.773946	
P(F<=f) one-tail	0.112191	
F Critical one-tail	2.182263	

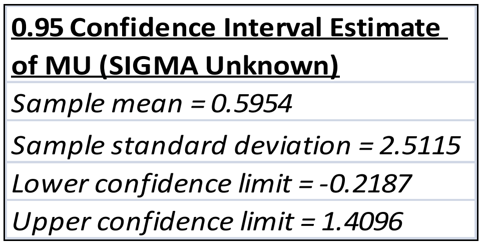
Table 3- Means and Variances



t-Test: Paired Two Sample for Means		
	Pretest	*Post test*
Mean	6.52673	7.07156
Variance	2.00119	2.35884
Observations	39	39
Pearson Correlation	0.36853	
Hypothesized Mean Difference	0	
df	38	
t Stat	-2.04856	
P(T<=t) one-tail	0.02373	
t Critical one-tail	1.68595	
P(T<=t) two-tail	0.04746	
t Critical two-tail	2.02439	

Table 4 - **Effect of Instruction**

Next, we checked if the results were uniformly positive. In the raw data, we noticed that the post test score for some students was actually lower than the pretest score. This was not related to corrections for missing data points or differences in ease of the two sets. We calculated the difference scores of the response pairs and determined the confidence interval at 95% confidence (**Table *5***). At this level, the confidence interval was partially located in negative territory. This meant that after going through the training sessions, some students performed worse rather than the same or better. We did try to relate this to the time spent on training, by regressing the time spent in training on the difference scores. The regression was not significant at p=.84, and time spent on training was therefore not a factor. In our opinion, negative difference scores are an indication that the participation of some students was not serious. Seen in this light, and considering that participation based on pure chance would result in a 5 point score on average, the actual learning effect may be somewhat understated.



0.95 Confidence Interval Estimate of MU (SIGMA Unknown)
Sample mean = 0.5954
Sample standard deviation = 2.5115
Lower confidence limit = -0.2187
Upper confidence limit = 1.4096

Table 5- Confidence Interval of Differences

4. CONCLUSIONS AND RECOMMENDATIONS

Based on the results of this quantitative study, some earlier studies are confirmed in a different setup. In particular, the findings of Kurumaguru et al (2007) indicate that instruction is much more effective when embedded. Though our training session was somewhat effective, the improved performance level would still be wholly inadequate to produce "safe" decisions in avoiding phishing emails. In our follow-up study, we plan to provide immediate feedback after each decision in the user training, and to increase the pool of legitimate and phishing emails so that users can train repeatedly with randomly selected examples. In the current design, the training set was static, though students could use it multiple times if they desired to do so. However, none of them did so the amount of training was limited. In contrast to Kurumaguru et al (2007; 2008), where the training was provided in the second half of the first session, we do intend to make the training session a separate event to more accurately separate training from performance. This is closer to the approach of Anandpara et al (2007), albeit that they provided the pretest, training, and post test in the same setting. Finally, the pretest and post test will have to be revised to make them more equal in performance. Whether we will do this by using alternating sets as in this study, or by creating different tests for pretest and post-test with similar characteristics, will be decided later. We do intend to limit the repeatability of the measurement, so that no improvement can be obtained just by taking the same test twice.

Our current study does have limitations. We use students, which is appropriate for an educational study, but not for generalization to the population at large. As such, the results could be very different in business environments. The sample size is adequate, especially considering the increased precision that repeated measures offer through elimination of interpersonal differences. This lends credibility to the validity of the results. Despite our best efforts, we were unable to produce and use two tests that demonstrably were interchangeable. Finally, as phishing attacks evolve, the materials for training and testing will need to be updated.

One of the drawbacks of some previous studies has been the use of customized email examples of legitimate and phishing emails. With our use of publicly available educational materials, and focusing on using these more effectively, we hope to contribute to increased email safety for the general public.

5. REFERENCES

Anandpara, V., Dingman, A., Jakobsson, M., Liu, D., and Roinestad, H., "Phishing IQ Tests Measure Fear, Not Ability," extended abstract, USEC, 2007.

Anti-Phishing Working Group. (2009). APWG Phishing trends reports. http://www.antiphishing.org/phishReportsArchive.html. Retrieved July 29, 2009.

Consumer Reports. (2009a, June). Boom time for cybercrime. Consumer Reports, 74, 18-21.

Consumer Reports. (2009b). "Phishing Trip." Retrieved July 29, 2009, from http://www.consumerreports.org/cro/electronics-computers/resource-center/cyber-insecurity/phishing-interactive.htm

Federal Deposit Insurance Corporation. (2008). "Phishing scams." Retrieved May 26, 2009, from http://www.fdic.gov/consumers/consumer/alerts/phishing.html

Federal Trade Commission. (2006). "How Not to Get Hooked by a 'Phishing' Scam." Retrieved July 29, 2009, from http://www.ftc.gov/bcp/edu/pubs/consumer/alerts/alt127.shtm

Federal Trade Commission. (2008). "Roundtable Discussion on Phishing Education", http://www2.ftc.gov/os/2008/07/080714phishingroundtable.pdf, accessed October 4, 2009

Jakobsson, M., & Myers, S. (2007). Phishing and Countermeasures: Wiley Interscience.

Kumaraguru, P., Rhee, Y., Sheng, S., Hasan, S., Acquisti, A., Cranor, L. F., et al. (2007). "Getting users to pay attention to anti-phishing education: evaluation of retention and transfer." Proceedings of APWG eCrime Researchers Summit, October 4-5. Pittsburgh, PA.

Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., & Hong, J. (2008). "Phishguru: lessons from a real world evaluation of anti-phishing training. " Proceedings of the eCrime Researchers Summit, October 15-16, Atlanta, GA.

Robila, S. A., & Ragucci, J. W. (2006). "Don't be a phish: steps in user education." Proceedings of the 11th Annual SIGCSE

Conference on Innovation and Technology in Computer Science Education, Bologna, Italy.

Sonicwall Inc. (2008). SonicWALL phishing and Spam IQ quiz. Retrieved May 26, 2009, from http://www.sonicwall.com/phishing/

The SANS Institute. (2009). SANS top-20. Retrieved July 29, 2009, from https://www2.sans.org/top20/#h2

W3counter. (2009). Global web stats. Retrieved July 2009, 2009, from http://www.w3counter.com/globalstats.php

Werner, L., & Courte, J. (2008, Nov 8). "Analysis of an anti-phishing lab activity." Proceedings of ISECON'08, Phoenix, AZ.