Evaluating Cybersecurity Vulnerabilities with the Email Testbed: Effects of Training

Ben D. Sawyer[1,2], Victor S. Finomore[2], Greg J. Funke[2], Vincent F. Mancuso[2], Brent Miller[2], Joel Warm[2,3] & P.A. Hancock[1]

[1]University of Central Florida, [2]USAF 711th Human Performance Wing, [3] University of Dayton Research Institute

Email-delivered cyber-attacks which penetrate first-line algorithmic defenses must then face a human operator's decision: engage, or 'reject and report' the threat. Relatively little research exists into the factors that affect these outcomes, and what does is contentious to conduct; naturalistic studies into cyber-attack can easily invade individual privacy. The development of our Email Testbed (ET), a simulator for clerical email tasks, is described and forwarded as an easy and ethical solution to these difficulties. In two experiments undergraduates first screened, then used the ET to send and receive workplace messages. They were concurrently delivered rare malicious emails. Half received cyber-defense training, and performed significantly better than their untrained peers. Without training, performance was poor enough so as to engender concern regarding real-world risks. Discussion of the process of validating the testbed, of the implications of the findings, and of future directions are provided.

Keywords: phishing, whaling, information security, cyber-defense

1. Introduction

The former Chief Scientist of the US Air Force (AF) described cyberspace as a domain through which all essential AF operations are performed (Maybury, 2012). In safeguarding this vital space, the focus of research and intervention has to date most often been on automated algorithmic systems and human teams of cyber-defenders (as in Finomore et al., 2013; Sawyer et al., 2014a). Such human-machine teaming is responsible for reducing the number and impact of attacks and identifying those missed, but crucially can never intercept all malicious messages. Human decisions, rendered by end users decide whether such attacks are successful. Email remains a primary form of communication for business and government, despite a growing field of competitors. As such, it represents the target of choice for attackers with the intent of delivering malicious code or harvesting valuable information, respectively referred to as 'malware' and 'phishing' attacks. Indeed, it is difficult to think of a military or commercial operation which does not in some fashion rely on email. The question is, when faced with an email-delivered cyber-attack, will the operator fail to detect and engage, or 'reject and report' the threat?

Concerns about both malware and phishing are well placed, as the incidence of both are growing. Phishing arose in the mid-1990s, while Happy99, the first widespread email virus, appeared in January of 1999 (Oldfield, 2001). Such email-delivered cyber-attacks have been growing in number now for a number of years. As of 2014 Symantec.cloud reported intercepting some 500,000 a day (Thonnard et al., 2012). The personal nature of email makes it ideal not only for generalized attacks, but also those directed at specific targets. This practice is termed 'spear-phishing', or when directed at high-value targets, is termed 'whaling'. Thonnard and colleagues reported that targeted email-delivered cyber-attacks were rarely intercepted in 2006, but by 2014 more than 100 a day were reported, with many more undoubted unreported or successful. As is the case for ordinance in many types of warfare, categories are increasingly blurred. Malware is used to phish and phishing tactics are used to deliver malware. An attack email can deliver a file, a link, or both. Either can lead to an improper request for information, the compromise of the system by malicious code, or, increasingly, both. By any name, email-delivered cyber-attacks are an issue of growing concern for commercial and government organizations (Mancuso et al., 2014a,b).

Given the decades that email-delivered cyber-attacks have existed (Oldfield, 2001), and their potential for damage, the relative paucity of literature addressing their rate of success in various contexts is surprising. A 2004 WestPoint study found 48% of those sent a targeted email improperly asking for personal information clicked the link (Coronges et al, 2012). Age may be a factor; in two studies younger participants were more likely to click on phishing links (Coronges et al, 2012; Sheng et al., 2010). Sex also influenced results; women

have been found more likely to click on phishing links than men, and also more likely to provide information after (Sheng et al., 2010). After-the-fact reports have found women 10% more likely to be victims (Jagatic, Johnson & Jacobson, 2007). There are also indications that level of education (Sheng et al., 2010) and field of study (Jagatic et al., 2007) played some role, with those with more education and studying science and math related fields clicking less often. Finally, training has been shown to reduce clicks on phishing links, with sex and age having no significant effect on efficacy of such training (Sheng et al., 2010). The lack of additional, or more detailed, accounts is best illustrated by this point: at present the most cited paper on the subject of phishing is an account of a class project at Indiana University (Jagatic et al., 2007).

This lack of hard numbers may represent the reluctance of institutions and companies to reveal the extent of their own vulnerability. On a personal scale, this same effect could explain why little information exists about how users protect their work and personal email accounts. Jagatic and colleagues (2007), who sent phishing emails to university students and employees, describe the considerable anger and suspicion with which their research was greeted. This backlash included calls for the closure of the research and firing of those responsible. A breach of privacy and trust within the confines of an experiment is, to the individual on the receiving end, a breach nonetheless. It must be acknowledged that naturalistic phishing and cyber-attack studies carry with them some risk of harm to participants, although seemingly mitigated by the educational benefits of being in such a study. Still, the scarcity of similar efforts may be explained by this issue. Finally, lack hard numbers may exist for the same reason that the ground truth of much of cyber-defense is unknown: a well-executed cyber-attack by definition may leave no trace (Sawyer et al, 2014b).

In the pursuit of understanding potentially risky applied situations, one commonly employed strategy is the use of simulation. Just as driving simulators allow researchers to observe dangerous roadway maneuvers (as in Sawyer et al., 2014a), simulation has been used in the cyber domain to explore cybervigilance (Sawyer et al., 2014b) without risk to a network and its users. Email lends itself to simulation, as standard email interface is readily reproduced and widely understood. At present no email-delivered cyber-attack simulator exists. In pursuit of that vision, in cooperation with the Air Force's 711th Human Performance Wing (711HPW), the Email Testbed (ET) was built. The testbed was designed to allow a participant to engage in a realistic email task for a prolonged period of time, while being attacked with malicious emails.

Two experiments were performed. The first, conducted online, piloted a training manipulation while allowing researchers to build a robust email corpus. The second used the tools developed in the first to run a laboratory experiment looking into the effects of training on detection of email-delivered cyber-attacks.

2. Experiment 1: Validation of the Email Corpus & Pilot Training Study

2.1 Description

The ET was designed to deliver a corpus of existing emails to a participant, allowing freedom to respond naturalistically at their own pace. The first challenge was to create such a corpus, capable of passing as as real and human generated (no small challenge: see Turing, 1951; 1952). To simplify this task, a narrow context was created, placing the participant in an administrative role for the fictitious company "Cog Industries". The participant's job: process forms containing sensitive information. Incoming emails would contain a request to either download or file a .pdf attachment, or upload and reply with an appropriate .pdf attachment. Legitimate emails had "cognind.com" as their email suffix, indicating they were internal to the company. Attack emails were distinguished by their email suffix, coming from various .tv domains. More than three hundred emails were written, along with a database of three hundred individuals and associated email addresses, salutations and signatures. It was expected that emails which sounded artificial would be flagged as 'suspicious', and as such false alarm (FA) percentage could be used as a measure of the corpus's suitability as seeming pass as 'real and human generated'. The plan was to identify emails with high FA percentage, and revise them in a series of iterations.

General hypotheses were formed. First, it was expected that the rewriting of high FA emails would drive overall FA percentage over time. Cyber-defense training was alternately withheld and added, with the expectation that its presence would improve hit percentage. Effects of sex and age, as described above, were further expected.

2.2 Method

Participants were recruited from the undergraduate population of a large southeastern university, and compensated with class credit. At the time we ran this email database validation, the ET was under

construction. Qualtrics.com, an online survey-making site, was used to present training and candidate emails. Qualtrics had no ability to mock up the uploading or downloading tasks, so participants were simply asked if each email should be reported or not. Participants completed the task online, in whatever setting they chose, and at their own pace.

Four iterations of this validation study were collected, each with changes to the underlying emails and training. In iteration I there was no training. A database of 327 emails was tested, of which eight were attack emails. Eight participants were recruited to iteration I, but one was dropped for non-completion (n = 7). Iteration II included the following elements: a) interface training, b) a description of the job the participant would be role-playing, c) descriptions of phishing and malware and d) seven examples of typical emails, four non-attack and three attack. A database of 330 emails was tested, of which sixteen were attack emails. Thirteen participants were recruited to this version, but two were dropped for non-completion (n = 11). In iteration III the following components were removed: c) descriptions of phishing and malware and d) seven examples of typical emails were removed. A database of 300 emails was tested, of which eight were attack emails. Sixteen participants were recruited to this iteration of which four were removed for non-completion (n = 12). In the final version, iteration IV, training was reinstated. A database of 300 emails was tested, of which eight were attack emails. Twenty-five participants were recruited to this version and two were removed for non-completion (n = 23). Given that age in all samples above ranged from 18 to 20, not unexpected for an undergraduate population, we did not have the range to detect age effects.

This signal probability chosen to be representative of the low percentage of cyber-attacks that evade first-line algorithmic defense was 1% (Sawyer et al., 2014b). Participants completed the task online, on their own computers. Ninety minute sessions were set aside for completion. Hit percentage (measured as attack emails reported over total attack emails), and FA percentage (measured as non-attack emails erroneously reported as attack emails over total non-attack emails), were collected.

2.3 Results

Table 1. Hit and False Alarm Percentage for Email Corpus Validation

Iteration	*n*	Hit %	SD	FA %	SD
I	7	19.0	18.8	22.2	16.0
II	11	83.7	26.0	6.1	8.0
III	12	56.3	38.9	3.8	5.9
IV	23	87.4	25.0	3.2	6.1

As can be seen by Table 1, the primary success of the first experiment was the improvement of the performance of the email corpus. Corpus changes and implementation of training between iterations I & II had large effects on both hit and FA percentages, driving the former up and the latter down. The removal of training in iteration III had a large impact on hit percentage, reducing it. Training was reinstated in iteration IV, and restored the lost hit percentage. Throughout this process emails responsible for FAs were being rewritten, but the attack emails were not changed between iteration three and four. The opportunity, therefore, was taken to statistically test the effectiveness of the simple cyber-defense training we had devised. A 2(training) x 2(sex) ANOVA was conducted on hit percentage data only, comparing iterations two & three. To compensate for the unbalanced sample size, a Type III sum of squares was used in the analysis. A main effect of training, $F(1,31) = 5.62$, $p = .02$, $\eta^2_p = 0.15$, was detected, but no interaction of sex by training or main effect of sex was seen.

Experiment 1 was a pilot conducted opportunistically amid construction of our testbed, and so necessarily suffered from considerable limitations: unequal sample size between conditions and confounds induced by revising the email corpus. Care was taken to mitigate these issues in Experiment 2.

3. **Experiment 2: Email Testbed Validation & Training Study**

3.1 Description

This testbed validation experiment used the validated email corpus and training from Experiment 1. It was run on the completed ET software, built in cooperation with the 711HPW. A purpose-build array of eight desktop computers with standard keyboards, mice, and 17 inch monitors displayed the testbed at a resolution of 1024x768. In order to minimize training, the ET interface was created to mimic popular webmail clients (Figure 1) with which research participants might already have experience. A mock file browser was included,

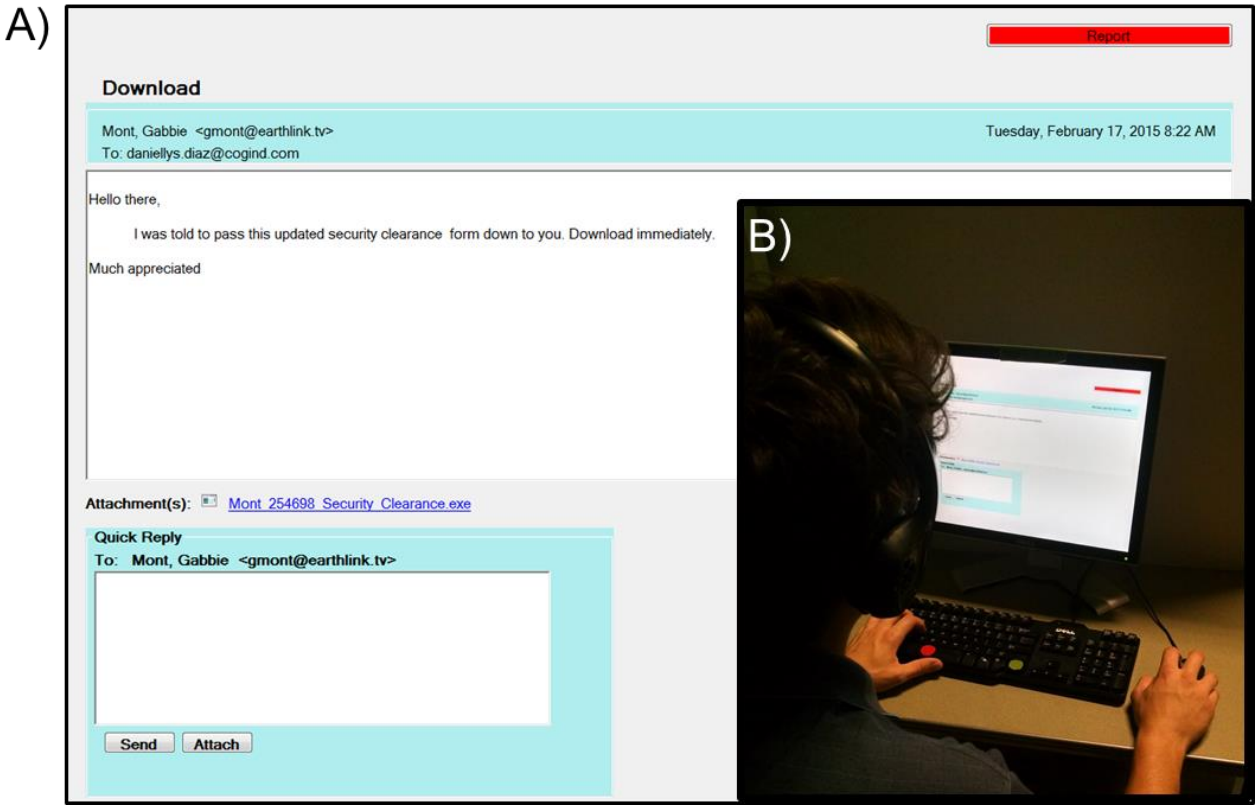

Figure 1. A) An attack email, as presented in the Email Testbed (ET). The interface of the ET was designed to mimic common online webmail interfaces, and to be minimal. In this email the inbound address, ending in .tv, the attachment, ending in .exe, and the body of the email, lacking a signature, all point to the suspicious nature of the email. B) (Inset) Participants at individual terminals interacted with the ET. Should a participant click on the executable, a miss would be recorded. Should the participant click on the red "Report" button, a hit would be recorded.

allowing participants to retrieve documents filed earlier, as well as draw from a list of pre-existing files. In this experiment non-attack emails originated from addresses in the cogind.com domain, and asked a participants to either download a file with the .pdf suffix and file it, or upload such an existing .pdf document. Attack emails all originated from non cognind.com domains, ended in .tv, and asked participants to download a file with a suffix ending in .exe. Of the three hundred emails provided, three were attack emails, for an overall signal probability of 1%. All participants received a description of the administrative position they would be role-playing in the experiment, as well as descriptions of the interface they were to use. If a participant detected a suspicious email, they were asked to push a red button marked "Report", which closed the email and returned the participant to the inbox. The manipulation was training based. Half of participants additionally received single PowerPoint slide with additional training directly related to email-delivered cyber-attacks (See Figure 2).

Figure 2. This basic cyber-defense training, presented as an image in Experiment 1 and a PowerPoint slide in Experiment 2, was the entirety of the training manipulation for both. Note that the definitions given fits both malware and phishing attempts. All were grouped under the term 'phishing' because participants in pilots expressed confusion as to whether malware could come from emails. As we note in the introduction, the definition of these two cyber-attacks increasingly overlaps.

3.2 Method

Fifty participants was recruited from the undergraduate population. Two were released for late arrival, leading to a final sex-balanced sample of forty eight (n = 48). Participants were compensated with class credit, and arrived at the laboratory in groups of eight or less. They were provided with informed consent and asked to turn off all cell phones or devices capable of creating an alert. Each was seated at a separate computer, and asked to view a PowerPoint slideshow providing interface and role descriptions. Participants in the 'training' condition received one slide of supplemental cyber-defense training (see above). They then began the experimental task, in which emails appeared serially; as soon as one was completed the participant was returned to the inbox, where another would appear. When all 300 emails had been interacted with, the program closed and a researcher guided the participant to a demographic portion. Participants were debriefed before leaving.

3.3 Results

Table 2. Hit and False Alarm Percentage for Email Testbed Validation

Training	Hit %	SD	FA %	SD
No	43.0	39.9	13.5	15.0
Yes	79.2	29.2	10.5	16.5

Visual inspection of Table 2 reveals an upward trend in hit percentage and downward trend in FA percentage for participants receiving training. Two 2(training) x 2(sex) ANOVAs were run on the dependant variables of hit percentage and FA percentage, each at a criterion of .025, for a familywise criterion of .05. In terms of hit percentage, there was a significant main effect of training, $F(1,44) = 12.78$, $p = 0.001$, $\eta 2p = 0.23$, in which training led to higher hit percentage. No significant sex by training interaction $F(1,44) = 0.68$, $p = .41$, $\eta 2p = 0.02$ or main effect of sex $F(1,44) = 1.21$, $p = 0.28$, $\eta 2p = 0.03$ were seen. In terms of FA percentage, no significant effects were detected.

4. **Discussion**

In both experiments, the effect of training was clear and significant, in line with the initial hypothesis. The number of untrained participants responding to (clicking on) the attack email, in this case by downloading an executable, was 57% in Experiment 2, as compared to 43% in the 2004 Westpoint study. Both numbers are alarmingly high. It is interesting to see just how effective the frankly minimal cyber-defense training was, which it should be recalled consisted of a single PowerPoint slide (Figure 2). In light of this finding, the cost of substantial workplace protection may be minimal, and the signs and computer pop-ups concerning safety that exist in so many institutional and commercial workplaces may have great utility.

Effects of age and sex were not seen. In the case of the former, the narrow range of our undergraduate sample precluded such detection. Trends for sex in our sample were in the direction identified by our hypothesis, but non-significant. The study did use the word "email" in the title participants saw when signing up, and so it is possible some self-selection occurred.

The development of a suitable email corpus can be considered a qualified success. The end false alarm percentage, 3.2 percent of non-attack emails being flagged as suspicious, was less than one sixth of the initial percentage. That said, given the signal probability of one percent used in Experiment 2, participants, on average, detected more signals that did not exist than did exist. Further, false alarm percentage among participants inflated from 3.6% in Experiment 1 to 10.5% in Experiment 2. This despite using the same corpus and training, was unexpected. This difference may be accounted for by demand characteristic differences between participants completing the experiments on their own machines in a setting of their choosing, and in the laboratory. It is further possible that there was a semantic difference between a binary suspicious/not suspicious question, and the need to press a button to flag an email as suspicious mid-task. It was considered possible that some individuals were using the single-click "report" button rather than the multiple clicks to download or upload as a quick way to move through the task. The researchers were able to make such a

cynical strategy work when trying it in the lab, however in the actual experiment those with higher FA did not show lower completion time.

It is difficult, as in so many cyber contexts, to know what "good" or "correct" looks like. Numbers that might clarify this question, such as the number of real emails that are accidentally misidentified by the human as malicious, are unavailable in the literature. This is, fortuitously, exactly the kind of question the ET is built to answer, and so now we add the question of such false-flagging to the category of 'future research which ET may address', along with collateral questions of user trust and acceptance.

In aggregate, these findings underline how vital the role of end user detection of cyber-attacks is, and how little is known about the factors that impact it. The Email Testbed, here presented for the first time, has already been able to shed some light. Future work with this tool should look to investigate training used in actual military and work settings, as well as to investigate attacks beyond downloadable executables. Attack emails in the present study were identified by email address and download extension, but future studies can look at each element of the interface and identify which signs of intrusion are most likely to be missed. It is hopeful that, with a simulator for clerical email tasks, selection and training strategies to reduce the success rate of email-delivered cyber-attacks can be devised.

Acknowledgements

MIT[2] Laboratory Research Assistants Rebecca McKeogh, Daniellys Diaz, James Wilcox and Isaac Yi were instrumental to building stimuli and collecting data for this project.

References

Coronges, K., Dodge, R., Mukina, C., Radwick, Z., Shevchik, J., & Rovira, E. (2012). The Influences of Social Networks on Phishing Vulnerability. Presented at the *2012 45th Hawaii International Conference on System Science*.

Finomore, V., Sitz, A., Blair, E., Rahill, K., Champion, M., Funke, G., ... & Knott, B. (2013). Effects of cyber disruption in a distributed team decision making task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57(1), 394-398.

Hong, J. (2012). The state of phishing attacks. *Communications of the ACM (Association for Computing Machinery)*,*55*(1), 74-81.

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social Phishing. *Communications of the ACM(Association for Computing Machinery)*, *50*(10), 94–100.

Mancuso, V. F., Strang, A. J., Funke, G. J., & Finomore, V. S. (2014a). Human factors of cyber attacks a framework for human-centered research. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), 437-441.

Mancuso, V. F., Christensen, J. C., Cowley, J., Finomore, V., Gonzalez, C., & Knott, B. (2014b). Human factors in cyber warfare II emerging perspectives. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), 415-418.

Maybury, M.T. (2012). *Cyber Vision 2025: United States Air Force Cyberspace Science and Technology Vision 2012-2025.* SAF/PA Public Release Case No. 2012-0439/460/715.

Oldfield, P. (2001). *Computer viruses demystified*. Sophos, *self published*.

Sawyer, B. D., Finomore, V. S., Calvo, A. A., & Hancock, P. A. (2014b). Google glass: a driver distraction cause or cure? *Human factors*, *56*(7), 1307-1321.

Sawyer, B. D., Finomore, V. S., Funke, G., & Warm, J. S. (2014a). Cyber vigilance: effects of signal probability and event rate. *Proceedings of the 2014 Human Factors and Ergonomics Society Annual Meeting*. 58(1), 1771-1775.

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L. F., & Downs, J. (2010). Who falls for phish?: A demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the Conference on Human Factors in Computing Systems* 13(1), 373–382.

Thonnard, O., Bilge, L., O'Gorman, G., Kiernan, S., & Lee, M. (2012). Industrial espionage and targeted attacks: Understanding the characteristics of an escalating threat. In *Proceedings of 15th International Symposium on Research in Attacks, Intrusions, and Defenses, 15(1),* 64-85.

Turing, A.M. (1950). Computing machinery and intelligence. *Mind*, 59, 433-460.

Turing, A.M (1951). Can digital computers think? In Furukawa,K., Michie, D., Muggleton, A. (Eds.), *Machine Intelligence, Vol. 15* (117-132). Oxford: Oxford University Press.