Human Factors in Cybersecurity, Vol. 127, 2024, 74–82
https://doi.org/10.54941/ahfe1004767

Does Penalty Help People Learn to Detect Phishing Emails?

Kuldeep Singh[1], Palvi Aggarwal[1], and Cleotilde Gonzalez[2]

[1]The University of Texas at El Paso, El Paso, TX 79968, USA
[2]Carnegie Mellon University, Pittsburgh, PA 15213, USA

ABSTRACT

Phishing attacks are increasingly prevalent and pose a significant threat to organizations worldwide. Many organizations implement phishing training programs to educate employees on how to recognize and avoid phishing attacks. Incentives are often used in these training programs to motivate employees to participate and engage with the material. However, the impact of incentives on the effectiveness of these training programs is not well understood. Similarly, how often such training should be provided, remains an additional factor in improving detection ability. Past research has provided evidence that frequency impacts the susceptibility to phishing emails. However, the interaction of frequency and incentives in phishing training is not well known. Key questions persist: Do individuals exhibit greater attention and motivation to detect phishing emails when penalties are imposed? How does exposure to more phishing emails contribute to evading penalties? This paper manipulates the frequency of phishing emails during the training phase and incentive structure for classifying emails. Experiments were conducted using a Phishing Training Task (PTT) i.e. an interactive software platform that emulates key tasks associated with email response decision making to test the impact of learning factors on phishing detection. The results indicate that imposing penalties for incorrect decisions does not have a significant effect on the detection performance for most of the conditions. Thus, our results suggest providing a symmetric incentive structure may not improve the phishing detection ability. These findings highlight the importance of experimenting with additional incentive structures in phishing training programs. This paper will provide guidelines to use cognitive models to design effective incentive structures.

Keywords: Phishing, Incentives, Phishing training, Cybersecurity

INTRODUCTION

Cyber attackers often use deceptive measure to exploit users. Phishing emails are among the most common form of deception, where attackers send fraudulent emails which resembles to a trusted email. Attackers use Cialdini's principles of persuasion that governs users to act according to the attacker's intensions (Cialdini 2001; 2004). According to Cialdini, attackers use reciprocation, consistency, social validation, liking, authority and scarcity to achieve positive response. Despite various efforts, phishing attacks continue to be rampant and successful. According to research from security software firm Trend Micro, over 91% of the cyberattacks start with a phishing attack (Trend Micro, 2023). The sophistication of phishing emails increased even more with generative AI algorithms (Sharma et al., 2023). Distinguishing phishing emails from benign remains a difficult task for people and falling prey to phishing emails has devastating consequences to both individuals and organizations (Parno et al., 2006). Attackers often take advantage of difficult and vulnerable situations while attempting phishing attacks. For example, during COVID, they use keywords such as "coronavirus," "COVID-19" and "Stimulus" to cheat people relying on unemployment benefits and stimulus payments (Symanovich, 2020). Automatic filtration of the phishing emails using machine learning algorithms is challenging. Therefore, training and awareness among individuals to detect phishing attacks is extremely important to improve cyber defense. Many organizations implement phishing training programs to educate employees on how to recognize and avoid phishing attacks. Incentives are often used in these training programs to motivate employees to participate and engage with the material. However, the impact of incentives on the effectiveness of these training programs is not well understood. Similarly, how often such training should be provided, remains an additional factor in improving detection ability. Past research has provided evidence that frequency impacts the susceptibility to phishing emails. However, the interaction of frequency and incentives in phishing training is not well known. Key questions persist: Do individuals exhibit greater attention and motivation to detect phishing emails when penalties are imposed? How does exposure to more phishing emails contribute to evading penalties?

In this paper, we frame an incentive structure that can have a significant effect on training and phishing detection. Incentive framing has been used in the context of structuring information to show positive or negative consequences of action (Levin et al., 1998). Incentives have shown an effect on the accuracy of detection in visual search tasks, although their effects may be delayed rather than immediate (Anderson et al., 2011; Fridrici et al., 2009; Madhavan et al., 2012). Rewards for correct classification and penalties for false alarms are examples of incentive framing. Such incentive-based training interventions are expected to lead users to focus on certain desired outcomes when performing a task (e.g., focusing on accurate classification). Another important factor to pay attention while designing training against phishing is the distractors in the training stimuli. While designing the training stimuli, Anderson et al. (2011) recommends avoiding the rewards for distractors in the training phase to avoid slowing down the learning process. In this paper, we test the effects of alternative incentive structures during training. Additionally, the employee's responses to such simulated phishing attacks are tracked and considered for annual bonuses, and thus incentivizing individual detection performance. Unfortunately, at the current moment, there is a lack of fundamental research that provides guidelines regarding what kind of incentives to provide to achieve high learning effects (Madhavan et al., 2012). To understand the role of incentives, we designed two experiments to test the effects of incentives on an individual's ability to learn to detect phishing attacks. We measured peoples' ability to discriminate phishing attacks from regular benign emails using Signal Detection Theory (SDT) measures. Overall, this work is intended to inform future anti-phishing training programs.

Literature Review

Phishing is a form of deception used by attacker to deceive the end-user and collect the sensitive information. The phishing attacks are getting sophisticated, because of attacker evolving techniques and strategies. Therefore, typical countermeasures struggle to counter these phishing attacks. Attackers target victim cognitive vulnerabilities using a diverse range of techniques. Anyone, lacking sufficient training or caught off guard, may fall victim to deception, potentially risking the entire organization (Singh et al., 2020). Training is indispensable for preparing individuals to avoid phishing attacks. Traditional approaches to phishing training involves delivering phishing educational material, classroom or slide-based training, and emails from security teams alerting employees about recent phishing attacks. These traditional methods of training typically prescribe people to take caution when clicking on links or downloading attachments from unknown sources. Regrettably, they fail to engage and motivate the users to learn the skills necessary for phishing detection (Anandpara et al., 2007). To address this limitation, organizations have turned to embedded training methods that involve sending simulated phishing emails and providing the more traditional phishing training. In simulated embedded training many factors such as frequency, recency, and feedback play a vital role (Singh et al., 2019; 2023). In addition to the type of training and feedback during training, incentives also play an important motivational factor for users (Goel et al., 2020; Jensen et al., 2020; McElwee et al., 2018). Security is often not the topmost priority of people, as day-to-day activities and deadlines takes over. Due to lack of awareness and poor security practices, enormous security breaches happen. Mustafa et al. (2021) study user behavior and their relationship with security management. To encourage security behavior, Herath et al. (2009) studied role of penalty, pressure, and perceived effectiveness. In order to motivate users to practice security habits, incentives may play a vital role. Maqbool et al. (2016) studied incentives as a motivational factor for cyber attackers and defenders. Muthal et al. (2017) studied the impact of incentive in email sorting task. The experiment results indicates that the participants took more time to process the emails with higher accuracy when provided incentive than otherwise. A study conducted by Zhang et al. (2018) suggest that, the monetary incentive can positively affect users' behavior and performance. There has been a limited work to study how incentives play a role in phishing training exercises. Goel et al. (2020) show that incentives could increase the overall compliance with security policies. Jensen et al. (2020) conducted experiments by manipulating the incentive structure as rewards only, punishment only, and both rewards and punishment. The results show that the punishment, even when joined with rewards, lowers the motivation, and reduces the number of hits (Jensen et al., 2020). The interaction of incentive structure with the training factors such as frequency is still unknown. This raises several questions about the design of embedded training that remain to be answered: How frequently should people be sent simulated attacks? How should end-users be incentivized? To answer these questions, one needs to analyse the cognitive processes involved in learning and decision making while processing emails. We address these questions using controlled experiments.

In this paper, we investigate how the frequency of phishing emails during training can improve phishing detection with different incentive structures. We expect the rewards along with punishment would improve the hit rate compared to using rewards only (Jensen et al., 2020). Singh et al. (2023) presented the results where the frequency of phishing emails was manipulated but no penalty was imposed for wrong classification. In this paper, we will be digging into the comparison between the outcomes laid out by Singh et al. (2023) and what happens when participants face penalties.

Methods

Using a Phishing Training Task (PTT) developed in Singh et al. (2019), we conducted an experiment to understand the role of incentives in phishing training. The PTT consists of three phases: pre-training, training, and post-training. The study implements a 2 × 3 factorial design to explore the impact of incentives and frequency on phishing susceptibility. The incentives were manipulated at two levels (No-Penalty and Penalty): no penalty for incorrect decisions (0 points) and penalty for incorrect decisions (-1 points). The rewards for correct classification were the same in both the conditions (1 point). The phishing email frequency was manipulated at three levels: low, medium, and high. In the low-frequency condition, 10 out of 40 emails were identified as phishing emails (25%). In the medium-frequency condition, 20 out of 40 emails were categorized as phishing emails (50%). Lastly, in the high-frequency condition, 30 out of 40 emails were classified as phishing emails (75%). Singh et al. (2019; 2023) presents results for the No-Penalty condition. In this paper, we present the interaction of incentives and frequency of phishing emails during training. We conducted two experiments, Penalty and No-Penalty, where participants were randomly assigned to one of the three experimental conditions: Low, Medium, and High. The details of the PTT task, email dataset and detailed procedure could be found in Singh et al. (2019; 2023).

Participants. All participants were recruited via Amazon Mechanical Turk (mTurk) and had a 90% or higher approval rate with at least 100 Human Intelligence Tasks (HITs) approved, resided in the United States. In the No-Penalty experiment, a total of 298 participants recruited and were randomly assigned to one of the 3 experimental conditions (low, medium, or high). Out of 298 participants, 2 participants failed both attention checks. The remaining 296 participants were distributed as follows: 98 participants in the low condition, 99 in the medium condition, and 99 in the high condition. In the Penalty experiment, we recruited 302 participants and randomly assigned to one of the 3 experimental conditions (Low, Medium, or High). Participants from No-Penalty experiment were not allowed to participate in this experiment. We checked the participant's attention in the experiment using two emails. Out of 302 participants, 5 participants failed in both attention checks. Thus, the remaining 297 participants were distributed as follows: 98 participants in low condition, 102 in medium condition, and 97 in high condition. On successful completion of the experiment, all participants were paid a base payment of $4 and up to $3 additional bonus, which was the same as No-Penalty experiment. The average time taken to complete this experiment was 30 minutes. We compare the participant's performance in Penalty experiment with No-Penalty experiment. Table 1 presents the demographics information of the participants.

Table 1. Demographics.

Experiment	Measure	Item	Value
No-Penalty	Gender	Male	57%
	Age	Mean	35.0
		SD	10.0
	Education	High school	15%
		Bachelor's Degree	37%
		Master's Degree	7%
		Some college education	38%
		Other	3%
Penalty	Gender	Male	57%
	Age	Mean	36.5
		SD	10.95
	Education	High school	14%
		Bachelor's Degree	42%
		Master's Degree	7%
		Some college education	35%
		Other	2%

RESULTS

We analyzed the change in sensitivity (Δd'), response bias (Δc) and over-precision of the participants in two experiments. The change in sensitivity, response bias and over-precision is the difference between post-training and pre-training phase performance.

The change in sensitivity and response bias is shown in Figure 1 and Figure 2. Over-precision for both the experiments is presented in Figure 3. The left panel of all figures represents the outcomes when participants faced no-penalty, while the right panel displays the results when a penalty was imposed. Table 2 summarizes the results of two-way ANOVA with frequency as between-subject factor.

Table 2. Two-way ANOVA with interaction effect.

DV	Factor	df	MSE	F-value	p-value
Δ d'	Experiment	1,587	3.61	0.01	0.913
	Frequency	2,587	3.61	1.26	0.284
	Experiment: Frequency	2,587	3.61	1.04	0.354
Δ c	Experiment	1,587	0.95	1.88	0.171
	Frequency	2,587	0.95	49.03	**<0.001**
	Experiment: Frequency	2,587	0.95	1.11	0.330
Δ Over-precision	Experiment	1,587	0.008	0.105	0.75
	Frequency	2,587	1.44	19.54	**<0.001**
	Experiment: Frequency	2,587	0.062	0.844	0.43

Sensitivity. The sensitivity results for the Penalty experiments in Figure 1 (right) show that the change in sensitivity is the high under low-frequency and high-frequency conditions. However, in the medium-frequency condition, the change is sensitivity is close to zero. According to the ANOVA in Table 1, there was no significant effect of frequency on the difference in sensitivity (Δd': F (2,587) = 1.26, p = 0.284). Thus, the frequency of phishing emails has no effect on the change in sensitivity. We compared the results between two experiments i.e. No-Penalty (Figure 1 left panel) and Penalty (Figure 1 right panel). We found no significant different between two experiments.

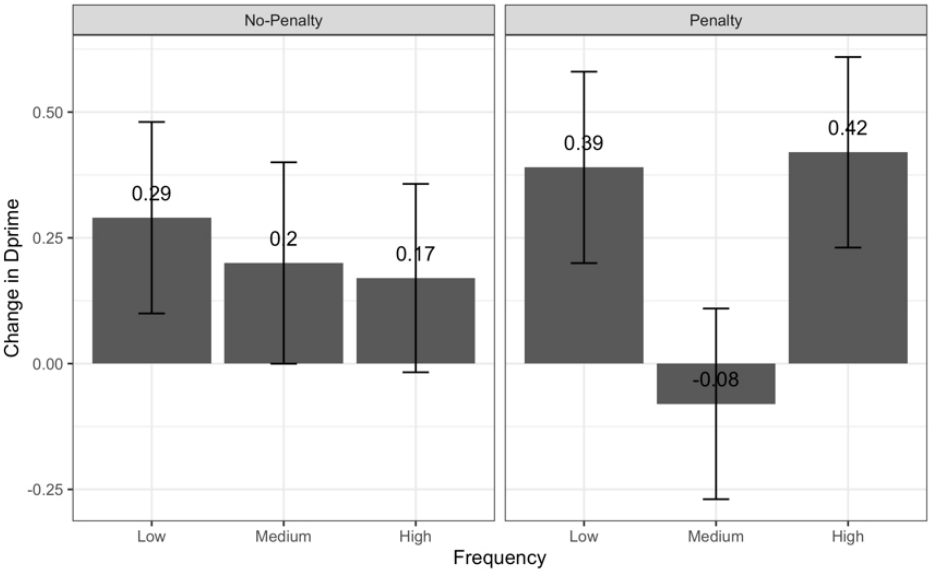

Figure 1: Change in Dprime.

Response Bias. Next, we analyze the change in response bias as shown in Figure 2 (right side). Response bias slightly increased in the low-frequency condition. However, the response bias decreased in the post-training phase compared to the pre-training phase under medium- and high-frequency conditions. Thus, the change in response bias is negative for both medium- and high-frequency conditions. Table 1 presents the significant difference in the change in response bias between the pre-training and post-training phases (F(2,587) = 49.03, p < 0.001). Post-hoc tests indicate that the change in response bias was higher in medium frequency compared to the low-frequency (*p.adj* < 0.001). Similarly the change in response bias is higher in high frequency compared to medium frequency conditions (*p.adj*<0.001and low frequency condition (*p.adj*<0.001). We compared the results between two experiments i.e. No-Penalty (Figure 2 left panel) and Penalty (Figure 2 right panel). We found that the change in response bias is similar in the presence and absence of penalty.

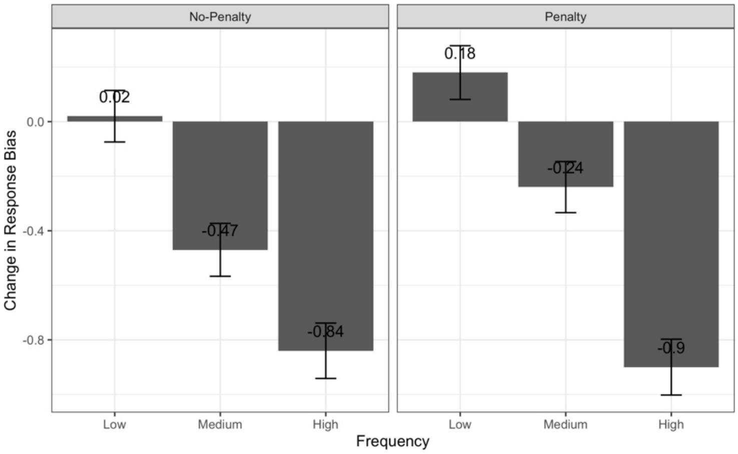

Figure 2: Change in response bias.

Over-precision. We illustrate the change in over-precision for the three frequency conditions in Figure 3 for both no-penalty (left panel) and penalty (right panel) conditions. In both no-penalty and penalty condition, the over-precision decreased after training in the low-frequency condition, that is, the confidence of the participants in their choices is lower after training compared to before training. We observe that the over-precision increases after training with medium and high-frequency conditions, indicating that participants were confident in their choices after the training. Table 1 indicates a significant effect of frequency (F(2, 587)= 19.57, p<0.001). Post hoc tests indicated that the change in over-precision was significantly different in all three frequency pairs ($p.adj$ <0.05). We also compared the two experiments, there was no significant difference between No-penalty and Penalty condition.

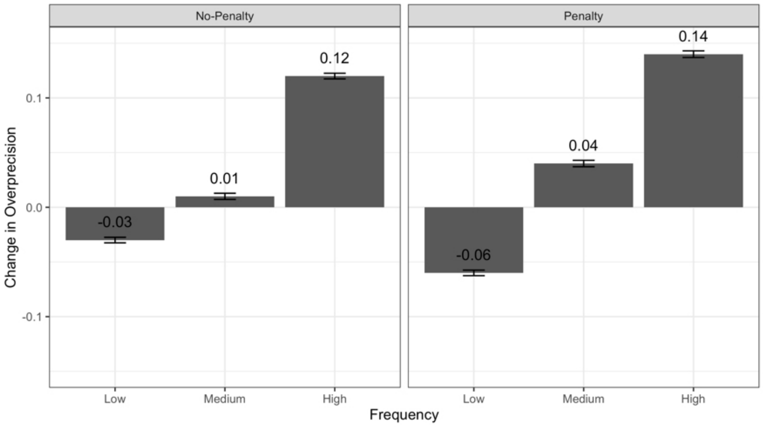

Figure 3: Change in over-precision.

CONCLUSION

In this paper, participant's performance in a phishing task was compared. In one experiment, the participants were given rewards for correct classification but were not penalized for incorrect responses; in another experiment, participants were penalized for incorrect responses. The results indicate that imposing penalties for incorrect decisions does not have a significant effect on the detection performance for most of the conditions. Research in Madhavan et al. (2012) suggests that different incentive structures may affect performance differently. In this paper, the incentive structure is symmetric, i.e., the rewards for correct classification and the penalty for incorrect classification are the same and termed as neutral incentives in Madhavan et al. (2012). This paper suggests that punishment for incorrect classification does not help in improving the phishing detection ability. The results presented in this paper confirms that the frequency of phishing emails in training phase impact the response bias (Singh et al., 2023). However, the sensitivity is not impacted by frequency or neutral incentive structure. In future, we plan to design different incentive structures to evaluate their impact on phishing training.

ACKNOWLEDGMENT

This research was sponsored by the Army Research Office and was performed under Grant number W911NF-17-1-0370 (MURI Cyber deception).

REFERENCES

Anandpara, V., Dingman, A., Jakobsson, M., Liu, D., & Roinestad, H. (2007). Phishing IQ tests measure fear, not ability. In Financial Cryptography and Data Security: 11th International Conference, FC 2007, and 1st International Workshop on Usable Security, USEC 2007, Scarborough, Trinidad and Tobago, February 12–16, 2007. Revised Selected Papers 11 (pp. 362–366). Springer Berlin Heidelberg.

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. Proceedings of the National Academy of Sciences, 108(25), 10367–10371.

Cialdini, R. (2001). Principles of persuasion. Arizona State University, eBrand Media Publication.

Cialdini, R. B. (2004). The science of persuasion. Scientific American Mind, 14(1), 70–77.

Fridrici, M., Lohaus, A., & Glass, C. (2009). Effects of incentives in web-based prevention for adolescents: Results of an exploratory field study. Psychology and Health, 24(6), 663–675.

Goel, S., Williams, K., Huang, J., & Warkentin, M. (2020). Understanding the role of incentives in security behavior.

Herath, T., & Rao, H. R. (2009). Encouraging information security behaviors in organizations: Role of penalties, pressures and perceived effectiveness. Decision Support Systems, 47(2), 154–165.

Jensen, M. L., Wright, R., Durcikova, A., & Karumbaiah, S. (2020). Building the Human Firewall: Combating Phishing through Collective Action of Individuals Using Leaderboards. Available at SSRN 3622322.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. Organizational behavior and human decision processes, 76(2), 149–188.

Madhavan, P., Lacson, F. C., Gonzalez, C., & Brennan, P. C. (2012). The role of incentive framing on training and transfer of learning in a visual threat detection task. Applied Cognitive Psychology, 26(2), 194–206.

Maqbool, Z., Aggarwal, P., Pammi, V. C., & Dutt, V. (2020). Cyber security: Effects of penalizing defenders in cyber-security games via experimentation and computational modeling. Frontiers in Psychology, 11, 11.

McElwee, S., Murphy, G., & Shelton, P. (2018, April). Influencing outcomes and behaviors in simulated phishing exercises. In SoutheastCon 2018 (pp. 1–6). IEEE.

Moustafa, A. A., Bello, A., & Maurushat, A. (2021). The role of user behaviour in improving cyber security management. Frontiers in Psychology, 12, 561011.

Muthal, S., Li, S., Huang, Y., Li, X., Dahbura, A., Bos, N., & Molinaro, K. (2017). A phishing study of user behavior with incentive and informed intervention.

Parno, B., Kuo, C., & Perrig, A. (2006). Phoolproof phishing prevention. In Financial Cryptography and Data Security: 10th International Conference, FC 2006 Anguilla, British West Indies, February 27-March 2, 2006 Revised Selected Papers 10 (pp. 1–19). Springer Berlin Heidelberg.

Sharma, M., Singh, K., Aggarwal, P., & Dutt, V. (2023, July). How well does GPT phish people? An investigation involving cognitive biases and feedback. In 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW) (pp. 451–457). IEEE.

Singh, K., Aggarwal, P., Rajivan, P., & Gonzalez, C. (2019, November). Training to detect phishing emails: Effects of the frequency of experienced phishing emails. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 63, No. 1, pp. 453–457). Sage CA: Los Angeles, CA: SAGE Publications.

Singh, K., Aggarwal, P., Rajivan, P., & Gonzalez, C. (2020, December). What makes phishing emails hard for humans to detect?. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 64, No. 1, pp. 431–435). Sage CA: Los Angeles, CA: SAGE Publications.

Singh, K., Aggarwal, P., Rajivan, P., & Gonzalez, C. (2023). Cognitive elements of learning and discriminability in anti-phishing training. Computers & Security, 127, 103105.

Symanovich, S. Coronavirus phishing emails: How to protect against COVID-19 scams.

Trend Micro Email threat landscape report: Cybercriminal tactics, techniques that organizations need to know. Security Roundup. (n.d.). https://www.trendmicro.com/vinfo/us/security/research-and-analysis/threat-reports/roundup/annual-trend-micro-email-threats-report.

Zhang, H., Singh, S., Li, X., Dahbura, A., & Xie, M. (2018, July). Multitasking and monetary incentive in a realistic phishing study. In Proceedings of the 32nd International BCS Human Computer Interaction Conference. BCS Learning & Development.