Scaling the Phish: Advancing the NIST Phish Scale

Fern Barrientos, Jody Jacobs(✉), and Shaneé Dawkins

National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
{fernando.barrientos,jody.jacobs,shanee.dawkins}@nist.gov

Abstract. Organizations use phishing training exercises to help employees defend against the phishing threats that get through automatic email filters, reducing potential compromise of information security and privacy for both the individual and their organization. These exercises use fake and realistic phishing emails to test employees' ability to detect the phish, resulting in click rates which the organization can then use to address and inform their cybersecurity training programs. However, click rates alone are unable to provide a holistic picture of why employees do or do not fall for phish emails. To this end, the National Institute of Standards and Technology (NIST) created the Phish Scale methodology for determining how difficult a phishing email is to detect [1]. Recent research on the Phish Scale has focused on improving the robustness of the method. This paper presents initial results of the ongoing developments of the Phish Scale, including work towards the repeatability and validity of the Phish Scale using operational phishing training exercise data. Also highlighted are the ongoing efforts to minimize the ambiguities and subjectivity of the Phish Scale, as well as the design of a study aimed at gauging the usability of the scale via testing with phishing exercise training implementers.

Keywords: Usable cybersecurity · Cybersecurity awareness training · Phishing · NIST Phish Scale

1 Introduction

Over half of all emails sent and received are spam; and an ever-growing number of those messages contain malicious threats [2]. Moreover, 10% of spam messages manage to get through email filters, and phishing emails account for approximately one-third of those emails [3]. Phishing emails are malicious threats designed to deceive and extract sensitive information from the email's recipient [4]. The phishing cyber threat exploits vulnerabilities in organizations of all types and sizes, including industry, academia, and government [5–8]. A major problem with phishing is that it targets what is possibly the most vulnerable element within any security system, the human user. While spam filters are capable of filtering phishing emails based on their sender, format, and verbiage, there are still phishes which get through this net, and it is these emails which can wreak havoc on an otherwise secured system. By clicking links and volunteering personally

C. Stephanidis et al. (Eds.): HCII 2021, CCIS 1420, pp. 383–390, 2021.
https://doi.org/10.1007/978-3-030-78642-7_52

identifiable information, those who have been successfully phished can end up costing themselves and their organizations a significant amount of money in recovery efforts and time lost.

To help combat the phishing threat, organizations strive to improve phishing awareness via embedded phishing training exercises. These exercises provide organizations data – click decisions in an operational environment – from realistic, safe, and controlled training experiences. However, these click decision data that show when an email user did or did not click on a link or attachment in a phish do not tell the whole story. The National Institute of Standards and Technology (NIST) Phish Scale (NPS) was conceived to provide context to these data – click rates – and to better understand why people do or do not fall for a phish [4]. The NPS is a method for determining how difficult or easy a phishing email is to detect [1] by considering both the characteristics of the email itself and the user context of the email's recipient. Ongoing research on the use of the NPS is intended to improve its robustness, validity, and ease of use. The goal of the research presented in this paper was to assess the repeatability and validity of the NPS when applied to phishing emails used during embedded phishing awareness training exercises.

2 Applying the Phish Scale

The Phish Scale was created to provide a metric for training implementers to gain a better understanding of the variability in click rates resulting from their phishing training exercises. The output of the NPS – a difficulty rating – can be used to provide context to these click rates. Steves, et al. previously described the NPS, its development, and its components in elaborate detail [1, 9]; a high level summary is presented below.

The NPS method is comprised of two major components. The first component is a measure of the observable characteristics, or cues, of the email itself (e.g., spelling, grammar). The more cues in a phish, the easier it is to detect. The second component, the premise alignment, measures how well an email aligns with the context of one's work. The higher the premise alignment, the more difficult the phish is to detect. For example, a phish that requests payment of an invoice is more difficult to detect (high premise alignment) to an individual in the accounts payable division. While the same invoice phish might be more easily to detect to a system architect whose job duties do not include payment of invoices. The NPS includes two separate approaches to determining the premise alignment – a *Formulaic Approach* and *Blended Perspective* [9]; the former approach is the focus of the analysis in this paper. When analyzed collectively, these two NPS components produce a difficulty rating for a target audience's susceptibility to a particular phishing email (Table 1). Phishing emails with a High premise alignment and Few cues are usually harder for individuals to detect. Conversely, emails with Low premise alignment and Many cues are easier to detect by individuals.

3 Research Methodology

One of the goals of this research presented in this paper is to gauge the repeatability of applying the NPS to phishing emails. To this end, the NPS was evaluated to measure the

Table 1. Determining detection difficulty

Number of cues	Premise alignment	Detection difficulty
Few (more difficult)	High	Very difficult
	Medium	Very difficult
	Low	Moderately difficult
Some	High	Very difficult
	Medium	Moderately difficult
	Low	Moderately to Least difficult
Many (less difficult)	High	Moderately difficult
	Medium	Moderately difficult
	Low	Least difficult

agreement between independent ratings of phishing exercise emails. This effort began with the process of reevaluating the phishing emails from a previously published paper on the NPS [9]. First, a team of NIST researchers (n = 3) who were not among the original authors of the NPS independently applied the NPS to the ten phishing emails originally published by Steves, et al. [9]. The team then met to assess and compare the individual scores for both cues and premise alignment. Points of divergence in cue counts or premise alignment element scores were discussed and ultimately resolved by averaging the scores of the team members. Finally, the team's consolidated scores and difficulty rating were evaluated against the previously published findings.

Another goal of this research is to validate the metric by applying the NPS to a broader set of phishing data. The first step toward this goal is presented in this paper; the NPS was applied to three additional phishing emails used in embedded phishing awareness training exercises throughout 2020 (see Appendix). The aforementioned steps were repeated in this effort – independent ratings by research team members followed by discussion and resolution of scoring conflicts. The results of these two research efforts are presented in Sect. 4.

4 Results

This section covers the results of our analysis of applying the NPS to the ten original phishing emails as well as the three additional phishing emails. As mentioned in the previous section, these 13 emails (ten original, three new) were independently rated by members of the research team and consolidated into final scores for the cue count and premise alignment. Each email was ultimately given an overall detection difficulty rating (referred to throughout the remainder of this section as the "new" scores and ratings).

For the original ten emails, the new scores and ratings were compared to the prior published work (see Sect. 3). This comparison is detailed in Table 2, where cue and premise alignment categories are specified for each phishing email, followed by the associated numerical score in parentheses. In addition to these data, the click rates (showing the percentage of email users who clicked in the email) for the phishing exercise associated with each phish are presented in Table 2.

When comparing cues, there is clear variance in the actual scores between the new and original analysis for the individual phishing emails. However, when abstracting up to their corresponding cue categories, agreement between the new categorical data and the original categorical data was met in 90% of these phishing emails. In regard to the premise alignment, the more subjective component of the NPS, an even greater variance is seen in the actual scores given by both new and original ratings. The effects of this numerical variance can be seen in the agreement between new categories and the original categories where ratings only matched up in 40% of the phishing emails. Lastly, in large part due to the variance in the results of applying the premise alignment component, the detection difficulty ratings were agreed upon in 50% of the ten original phishing emails when comparing the original data to the new data. Given the five possible ratings on the scale of detection difficulty, it is important to note that while 50% of the new ratings did not match the original ratings, the differences were only by a factor of one (e.g., "very" to "moderately" rather than "very" to "least"). Additionally, when comparing the original detection difficulty ratings and click rates to the new ratings and corresponding click rates, the new ratings exhibit a similar pattern to the original ratings in how they line up with the click rates.

Table 2. Comparison of NIST Phish scale ratings for original phish emails

Phish email	Cues (new)	Cues (original)	Premise alignment (new)	Premise alignment (original)	Difficulty (new)	Difficulty (original)	Click rates
E1	Few (6)	Few (7)	Low (10)	High (30)	Moderate	Very	49.3%
E2	Some (10)	Some (14)	Medium (13)	High (24)	Moderately	Very	43.8%
E3	Few (7)	Few (8)	Medium (16)	High (24)	Very	Very	20.5%
E4	Some (9)	Few (6)	Medium (14)	High (18)	Moderately	Very	19.4%
E5	Some (9)	Some (11)	Low (9)	Medium (14)	Moderately to least	Moderately	11.6%
E6	Some (13)	Some (13)	Low (0)	Low (10)	Moderately to least	Moderately to least	11.0%
E7	Many (18)	Many (18)	Medium (13)	Medium (16)	Moderately	Moderately	9.1%
E8	Some (9)	Some (12)	Medium (12)	Medium (12)	Moderately	Moderately	8.7%
E9	Some (14)	Some (11)	Low (−1)	Low (2)	Moderately to least	Moderately to least	4.8%
E10	Some (10)	Some (12)	Medium (13)	Low (4)	Moderately	Moderately to least	3.2%

Table 3 features the averaged calculations for the three independent raters of the current study. The click rates for emails E11 and E12 align well with their respective detection difficulty ratings, according to the pattern exhibited in the application of the NPS to the previous ten emails. However, the click rates for email E13 do not fully align with the established detection difficulty rating scale. The trend of the NPS has been for emails with a click rate as low as 2.8% to have a "least" or "moderately to least" difficulty rating.

Table 3. NIST Phish Scale ratings for new phish emails

Phish emails	Cues	Premise alignment	Detection difficulty	Click rates
E11	Some (12)	Low (4)	Moderately to least	12.7%
E12	Many (18)	Low (9)	Least	5.4%
E13	Many (16)	Medium (13)	Moderately	2.8%

The NPS has the ability to contextualize click rates with its detection difficulty ratings. However, there are some unexpected factors which may inflate click rates which could lead to the disagreement between click rates and detection difficulty ratings (as exhibited by E13). For example, when a phishing email appeared to come directly from an authority figure in upper management, it elicited serious concerns and a deeper sense of action by the email recipient than was measurable by the NPS, ultimately leading to an increased click rate and the aforementioned disagreement. These factors are intended to be addressed in future iterations of NPS development.

5 Discussion and Future Work

This paper presents an initial look into the ongoing validation effort of the NPS. The results discussed in the previous section show the margin of error in the NPS difficulty rating determination; there can be a slight variance in independent scores of a phishing email, yet that variance is not reflected in the resulting detection difficulty rating. This provides insight into the development of future iterations of the NPS; however, additional validation testing is needed, including testing with larger and more diverse datasets. To this end, the NPS is currently being tested with a variety of large datasets (both public and nonpublic) from universities, private companies, and other government agencies. The findings from applying the NPS to a variety of datasets will be used to improve future iterations of the NPS. These efforts are aimed at ensuring the NPS's accuracy and validity.

NIST is conducting a research study to determine the usability and applicability of the NPS. The study invited both federal and non-federal organizations with robust phishing programs to apply the NPS in their organizations, aligning with their existing embedded phishing awareness training programs. Following their use of the NPS, training implementers were asked to provide detailed feedback and recommendations about their use of the NPS. This valuable real-world information resulting from the study will determine the effectiveness of the NPS in unique organizational environments, how usable the NPS is, and how organizations use the NPS to contextualize phishing exercise click rates.

As mentioned throughout this paper, NPS research is ongoing. Current efforts to improve repeatability, to evaluate validity, and to assess the usability of the NPS are expected to lead to a more streamlined version of the NPS that would be beneficial to organizations to provide clarity, functionality, and adaptability of the metric. Future iterations of the NPS will incorporate various modifications grounded in findings from the research. Revisions currently being considered for adoption and inclusion are: 1) for the observable cues component, reducing subjectivity, increasing identification accuracy, and minimizing redundancy across the scale, 2) refining the cue counting method by incorporating a weighting metric to address cue saliency, and 3) restructuring the premise alignment's five elements to be more efficient, reducing the total number of elements and adopting proven methodologies for determination of premise alignment element scores. Additionally, insights gleaned from the aforementioned usability study, including the identification of successful practices and strategies, and lessons learned will be used to refine future iterations of the NPS.

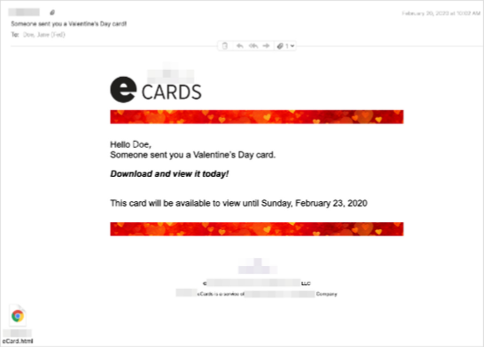
6 Conclusion

The NPS helps organizations and phishing awareness training implementors in two primary ways. Firstly, by contextualizing message click and reporting rates for a target audience, and secondly by providing a way to characterize actual phishing threats so training implementors can reduce the organization's security risk. Organizations should tailor their cybersecurity and privacy awareness training program to their unique environment while still meeting their organizations' mission and risk tolerance. Likewise, the NPS goes beyond the face value of an email by accounting for the environment, roles, and responsibilities of people within an organization. Tailoring training to the types of threats their organization faces helps them maintain a resilient security and privacy posture. Additionally, when click rates and quantitative and qualitative metrics from the NPS are viewed holistically, they can signal to an organization that training approaches and objectives, delivery methods, training frequency or content necessitate alterations to be effective in combating the ever-changing phishing threat landscape.

Disclaimer. Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Appendix

Note: Logos have been blinded from the phishing email images below (Figs. 1, 2 and 3).



[redacted]

Someone sent you a Valentine's Day card!

To: Doe, Jane (Fed)

February 20, 2020 at 10:02 AM

[redacted]

e CARDS

Hello Doe,
Someone sent you a Valentine's Day card.

Download and view it today!

This card will be available to view until Sunday, February 23, 2020

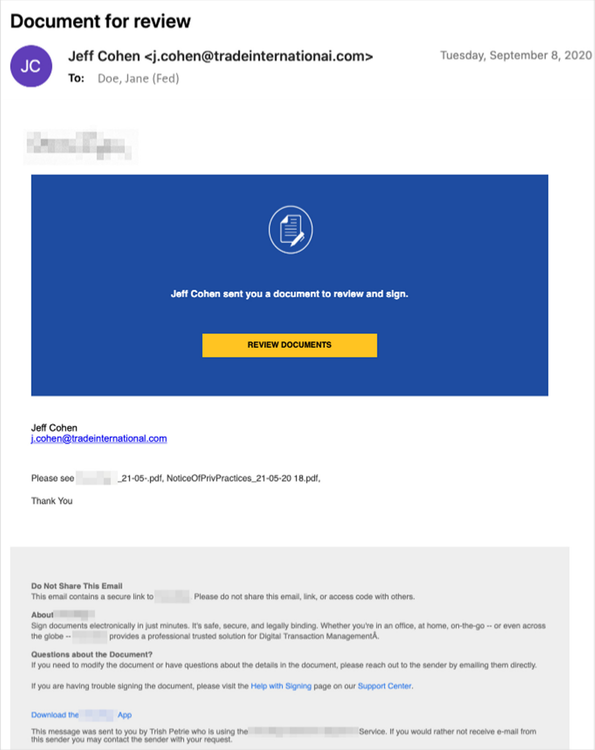
[redacted]

© [redacted] LLC

[redacted] eCards is a service of [redacted] Company

[redacted]
eCard.html

Fig. 1. E11: E-card phish



Document for review

JC Jeff Cohen <j.cohen@tradeinternationai.com>

Tuesday, September 8, 2020

To: Doe, Jane (Fed)

[redacted]

Jeff Cohen sent you a document to review and sign.

Jeff Cohen
j.cohen@tradeinternational.com

Please see [redacted] _21-05-.pdf, NoticeOfPrivPractices_21-05-20 18.pdf,

Thank You

Do Not Share This Email
This email contains a secure link to [redacted]. Please do not share this email, link, or access code with others.

About [redacted]
Sign documents electronically in just minutes. It's safe, secure, and legally binding. Whether you're in an office, at home, on-the-go -- or even across the globe -- [redacted] provides a professional trusted solution for Digital Transaction ManagementÂ.

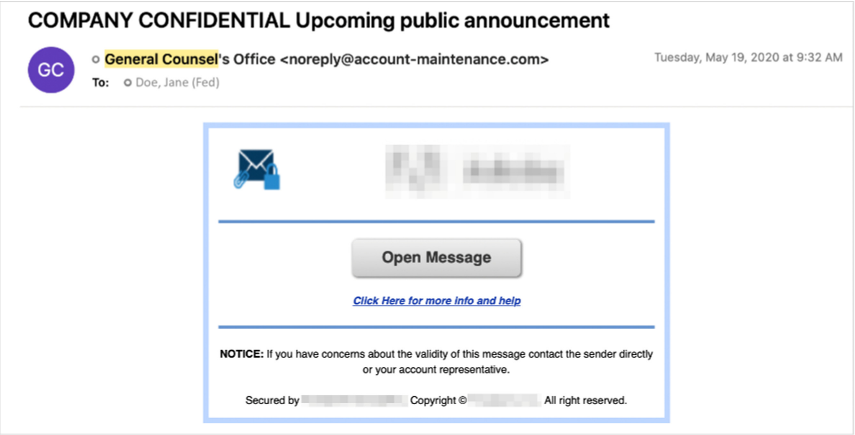
Questions about the Document?
If you need to modify the document or have questions about the details in the document, please reach out to the sender by emailing them directly.

If you are having trouble signing the document, please visit the Help with Signing page on our Support Center.

Download the [redacted] App

This message was sent to you by Trish Petrie who is using the [redacted] Service. If you would rather not receive e-mail from this sender you may contact the sender with your request.

Fig. 2. E12: Document review phish



COMPANY CONFIDENTIAL Upcoming public announcement

○ General Counsel's Office <noreply@account-maintenance.com> Tuesday, May 19, 2020 at 9:32 AM

To: ○ Doe, Jane (Fed)

NOTICE: If you have concerns about the validity of this message contact the sender directly or your account representative.

Secured by [redacted] Copyright © [redacted] All right reserved.

Fig. 3. E13: Public Announcement phish

References

1. Steves, M.P., Greene, K.K., Theofanos, M.F.: A phish scale: rating human phishing message detection difficulty. In: Proceedings 2019 Workshop on Usable Security. Workshop on Usable Security, San Diego, CA. (2019). https://doi.org/10.14722/usec.2019.23028
2. Ezpeleta, E., Velez de Mendizabal, I., Hidalgo, J. M.G., Zurutuza, U.: Novel email spam detection method using sentiment analysis and personality recognition. Logic J. IGPL, **28**(1), 83–94 (2020). https://doi.org/10.1093/jigpal/jzz073
3. Cyren: Email Security Gap Analysis: Aggregated Results. (2017). https://pages.cyren.com/rs/944-PGO-076/images/Cyren_Report_GapAnalysisAgg_201711.pdf. Accessed Jan 2021
4. Greene, K.K., Steves, M., Theofanos, M., Kostick, J.: User context: an explanatory variable in phishing susceptibility. In: Proceedings of 2018 Workshop Usable Security (USEC) at the Network and Distributed Systems Security (NDSS) Symposium (2018)
5. Aaron, G., Rasmussen, R.: Global phishing survey 2016: Trends and domain name use. Anti-Phishing Working Group. June 2017. https://docs.apwg.org/reports/APWG_Global_Phishing_Report_2015-2016.pdf. Accessed Aug 2017
6. Hong, J.: The state of phishing attacks. Commun. ACM **55**(1), 74–81 (2012)
7. SonicWall: 2019 SonicWall Cyber Threat Report (2019). https://www.sonicwall.com/resources/white-papers/2019-sonicwall-cyber-threat-report. Accessed Aug 2020
8. Symantec: Internet Security Threat Report. vol. 24 (2019). https://www.symantec.com/security-center/threat-report. Accessed Aug 2020
9. Steves, M., Greene, K., Theofanos, M.: Categorizing human phishing difficulty: a Phish scale. J. Cybersec. **6**(1), tyaa009 (2020). https://doi.org/10.1093/cybsec/tyaa009