Improving Phishing Risk Perception with Transparent Suspicion Scores

Peifeng He, Kejia Liu, Ethan Huang, Pardis Emami-Naeini

Duke University

Introduction

Email remains a widely used tool due to the ease it offers in communication, but this also makes it a prime target for phishing attacks. Various phishing email-detecting methods have been developed. Among those methods, suspicion scores are particularly common, offering users a numerical indicator of how suspicious an email might be [1, 3]. However, suspicion scores have notable limitations. They present a single value without context, making them appear as opaque "black-box" outputs. This lack of transparency prevents users from understanding which parts of the email are suspicious. In addition, people could interpret the same score differently. **These drawbacks can lead to inaccurate assessments and introduce high variance in risk judgments.**

To address these issues, we enhance suspicion scores with transparent, explanatory feedback. The goal is to inform users which specific factors—like sender mismatch, urgent language, or suspicious links—contributed to the score. A user study with 31 participants evaluated this approach. **Results showed that participants were better at identifying the actual risk factors and showed less variability when explanations were provided.**

User Study

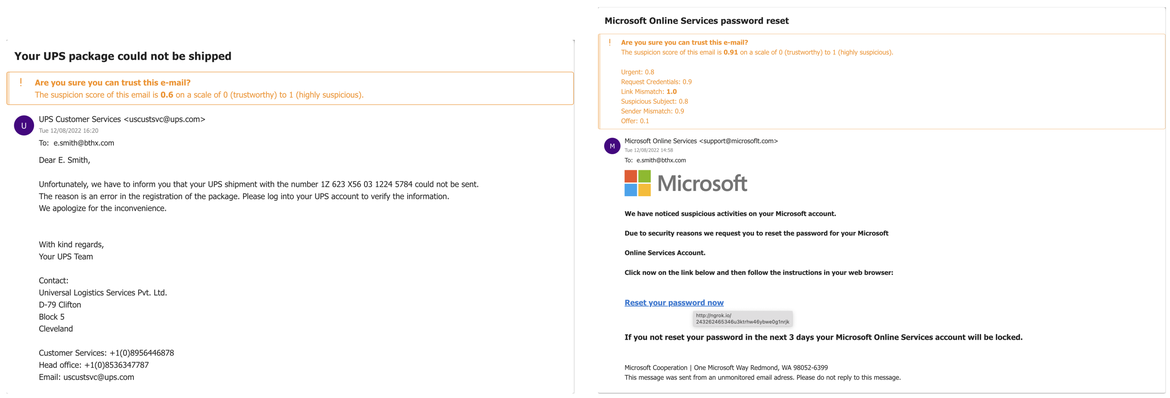
We conducted an online user study via a survey. In order to simulate real life or work scenarios, we ask participants to handle five emails as E. Smith, an employee of a virtual company called "BTHX".

Sender	Phishing Characteristics	Score
j.bieber@bthx.com	Not a phishing e-mail	0.22
m.greggs@tucsonweekend.com	Suspicious subject, offer	0.90
support@microsoflt.com	Link mismatch, request credentials, urgent	0.91
uscustsvc@ups.com	Sender mismatch	0.60
m.locke@bthx.com	Link mismatch	0.75

Table 1. Phishing characteristics and suspicion scores of five surveyed e-mails.

Our emails were presented to participants using a mock inbox that simulated the Outlook UI[3]. The study followed a two-phase design.

- Phase 1: participants viewed five email screenshots with suspicion scores only. They rated the likelihood of each being a phishing attempt on a 10-point scale and noted the factors behind their judgment. They also indicated the score threshold above which they considered an email risky.
- Phase 2: the same emails were shown again but with transparent explanations[2] of the suspicion scores. Participants re-rated each email and indicated whether the explanations helped them assess the risk.



(a) Suspicion score *only* (b) Suspicion score *with explanation*

The left image shows what the emails looked like in Phase 1, and the right image shows the emails presented in Phase 2.

Results

We found that when there is only one score, participants have limited ability to identify whether an email is risky (figure 1).

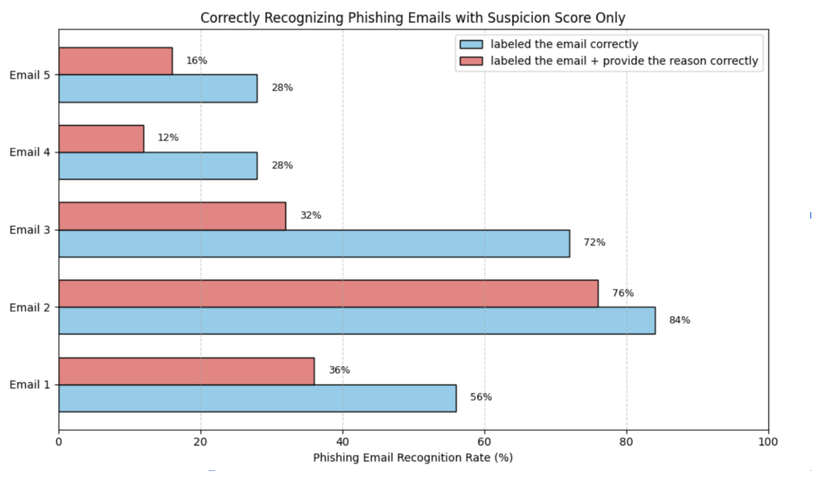


Figure 1. The percentage of emails and phishing reasons correctly identified in *phase 1*

Because different people have different understandings of the score, participants have very different views on the potential dangers represented by the same score (figure 2).

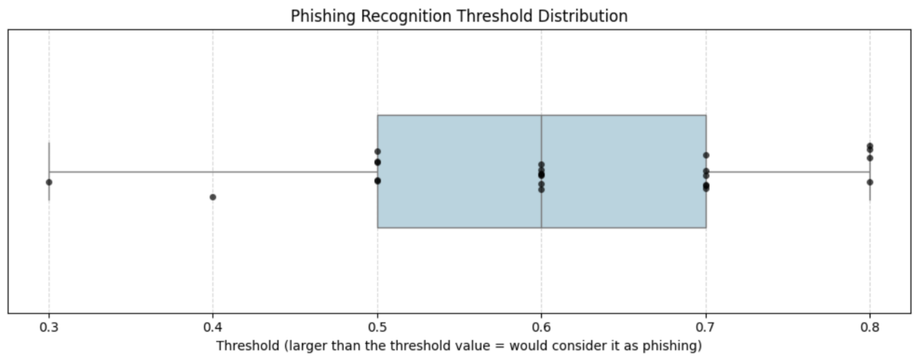


Figure 2. Distribution of risk threshold

As displayed in figure 3, for each email, when there is a transparent explanation, the participants' perception of the email risk factor is closer to the actual risk factor of the email.

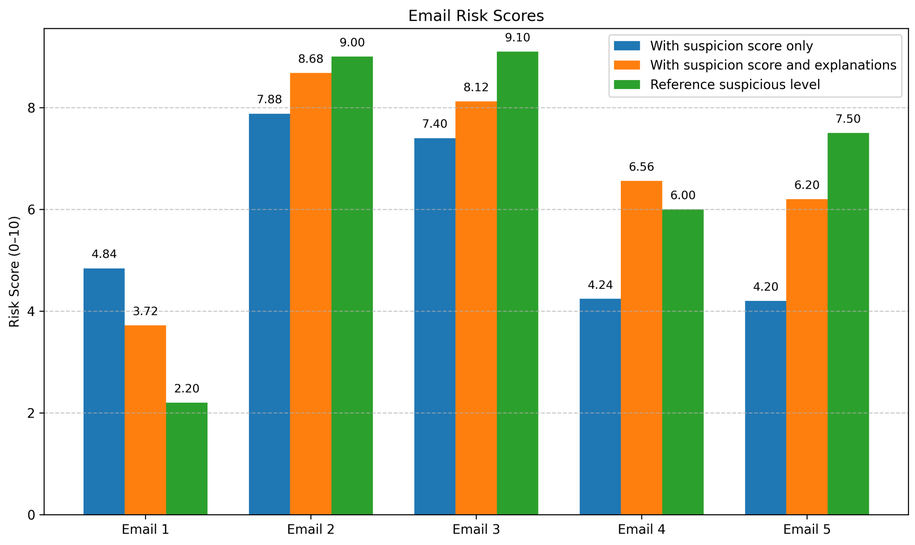


Figure 3. Average of scores before and after detailed explanations are added for 5 emails, comparing with the ground truth score

As shown in figure 4, except for the third email where the STD increased, the STDs of the other four emails decreased. We believe that the reason for the increase in STD in the third email may be that some participants did not notice the links introduced by the curser. In general, due to the decrease in STD, participants' risk perception of emails became more unified and they were more inclined to reach a consensus.

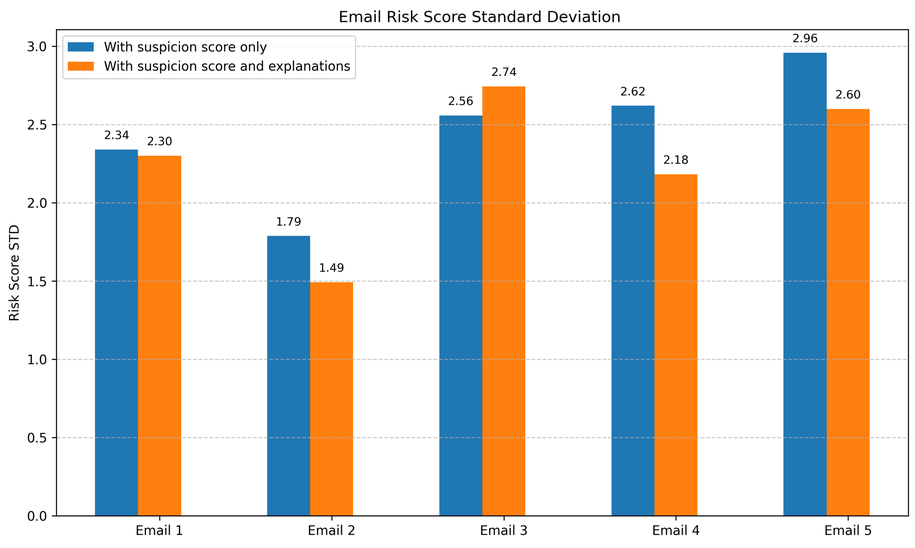


Figure 4. Standard deviation of scores before and after detailed explanations are added for 5 emails

Discussion and Conclusion

We discovered that providing explanatory information further enhanced users' judgment accuracy and reduced individual variation. This contrasts with earlier studies that treated suspicion scores as standalone indicators. We attribute this improvement to the added transparency, which helps users better understand the rationale behind the score, thereby building trust and supporting more informed decisions. This suggests that explanation-aware design can significantly improve the usability and effectiveness of phishing detection systems.

References

[1] Giuseppe Desolda, Francesco Di Nocera, Lauren Ferro, Rosa Lanzilotti, Piero Maggi, and Andrea Marrella. *Alerting Users About Phishing Attacks*, pages 134–148. 06 2019.

[2] Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. What makes phishing emails hard for humans to detect? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 64(1):431–435, 2020.

[3] Sarah Y. Zheng and Ingolf Becker. Checking, nudging or scoring? evaluating e-mail user security tools. In *USENIX Symposium on Usable Privacy and Security (SOUPS)*, pages 57–76, 2023. `https://www.usenix.org/conference/soups2023/presentation/zheng`.