SECURE MENTEM

WHITEPAPER

A user study of warning banners as a real-time threat assistant to reduce clicks on malicious emails.

Secure Mentem analysis of INKY Phish Fence

Researched and Written by:

Ira Winkler, CISSP
Secure Mentem
securementem.com

Introduction

While the usefulness of INKY is intuitively obvious, INKY wanted to have a definitive and scientifically sound study performed to statistically prove what anecdotal evidence implied. INKY retained Secure Mentem to perform a study with scientific rigor that examined how effective INKY banners were in influencing user determination of the safety of a given email message.

Secure Mentem designed a study that had 500 email messages, where approximately 100 messages were phishing messages collected from the Internet and 400 messages were safe. We then had the same messages with the INKY banners embedded within the messages. Secure Mentem then had 100 users evaluate messages with the banners (the test group) and another 100 users evaluate the same messages without the banners (the control group).

The study proved that the users were significantly more likely to correctly identify phishing messages. Users were 48% more likely to identify the most risky phishing messages, which would significantly reduce phishing risk in organizations. As important, the study found that users were able to more accurately identify safe email messages with the banners compared to without the banners, which should increase the operational efficiency in the organization.

Method

A total of 200 participants were recruited from an online research platform, with 100 participants assigned to view emails displaying banners (Banner Group) and 100 assigned to view emails without banners present (No Banner Group). Participants were instructed to indicate whether an email was “safe” or “unsafe” by clicking a designated button and confirming their decision.

Participants in the Banner Group viewed and made judgments on a total of 20 potentially unsafe emails with 16 displaying a YELLOW Banner, 4 potentially unsafe messages displaying a RED Banner. A total of 80 safe messages which displayed a GREY Banner were also presented. Participants in the No Banner Group viewed and made judgments on the same email messages but without the banners displayed. All email messages were displayed in a randomized order.

Email messages were collected from a honeypot domain set up on the Internet. We seeded the Internet to have messages sent to the domain which also began to attract spam and phishing emails. Secure Mentem worked with INKY to cull the messages for safe and phishing messages. These were the messages used for the study.

Results

Participants in the Banner Group were 13% more likely to correctly identify potentially malicious emails (78% Correct Identification) than the No Banner Group (65% Correct Identification). This improved performance was statistically significant[1] at less than a .001% probability of false positive[2], [t (198) = 3.962, p<.001], with a moderate effect size[3] [Cohen's d = 0.56].

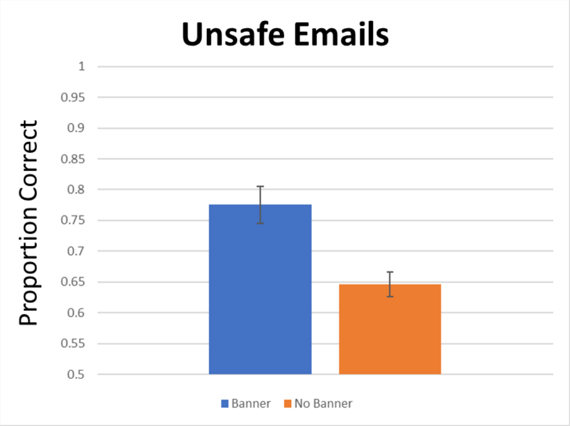

Figure 1: Accuracy on Unsafe Email Detection

[1] https://en.wikipedia.org/wiki/Student%27s_t-test
[2] https://en.wikipedia.org/wiki/P-value
[3] A commonly used interpretation is to refer to effect sizes as small (d = 0.2), medium (d = 0.5), and large (d = 0.8) based on benchmarks suggested by Cohen (1988).

Unsafe Emails: Accuracy			
Category	N	Mean	Std. Error
Banner	100	.7754	.02534
No Banner	100	.6460	.02061
t (198) = 3.962, p<.001, d = 0.56			

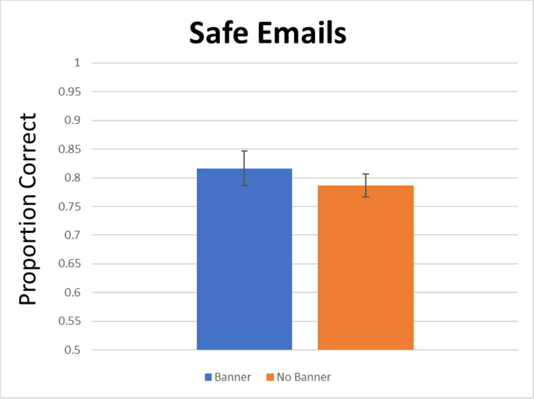

Figure 2: Accuracy on Safe Emails

Safe Emails: Accuracy			
Category	N	Mean	Std. Error
Banner	100	.8165	.02125
No Banner	100	.7865	.01935
t (198) = 1.044, p=.298			

Interesting results emerged when comparing the Red and Yellow Banners. The Banner Group were 26% more accurate than the No Banner Group when viewing emails with Red Banners (highest risk messages) (p<.001, d=.934) with a large effect size[3]. This is a 48% increase in risk reduction and accuracy. The Banner Group was 11% more accurate when viewing Yellow Banner emails (p<.001, d=.49), this represented a moderate improvement[3].

<table>
<tr><th colspan="4">RED Banners: Accuracy</th></tr>
<tr><th>Category</th><th>N</th><th>Mean</th><th>Std. Error</th></tr>
<tr><td>Banner</td><td>100</td><td>.8025</td><td>.02979</td></tr>
<tr><td>No Banner</td><td>100</td><td>.5400</td><td>.02629</td></tr>
<tr><td colspan="4">t (198) = 6.607, p<.001, .934</td></tr>
</table>

<table>
<tr><th colspan="4">YELLOW Banners: Accuracy</th></tr>
<tr><th>Category</th><th>N</th><th>Mean</th><th>Std. Error</th></tr>
<tr><td>Banner</td><td>100</td><td>.7840</td><td>.02569</td></tr>
<tr><td>No Banner</td><td>100</td><td>.6707</td><td>.02030</td></tr>
<tr><td colspan="4">t (198) = 3.461, p<.001, d=.49</td></tr>
</table>

Signal Detection Analysis: Studies involving judgement must exercise caution against relying solely on proportion correct or incorrect when measuring performance because this metric does not account for bias in decision makers. For example, in this study, if a participant simply answered unsafe on every email, they would have a 100% accuracy for the unsafe emails, but this would simply indicate a heavily biased decision maker. To better understand participant performance, the ratio of correct identifications (hits) to false positives (false alarms) should be considered. When considering these ratios two metrics of performance become relevant, the bias (λ), and the ability to discriminate between unsafe and safe emails (d’). The difference in d’ indicates that the Banner Group was better able to discriminate between safe and unsafe emails, the small λ indicates that both groups experienced a similar bias (proclivity) toward judging an email as safe or unsafe. This finding reflects positively on the influence of email banners because it demonstrates that neither group had a proclivity toward answering in one direction (either biased toward answering safe or unsafe), but that the Banner Group was more effective in judging which emails were potentially safe or unsafe. This suggests that the banners are helping users make better judgements.

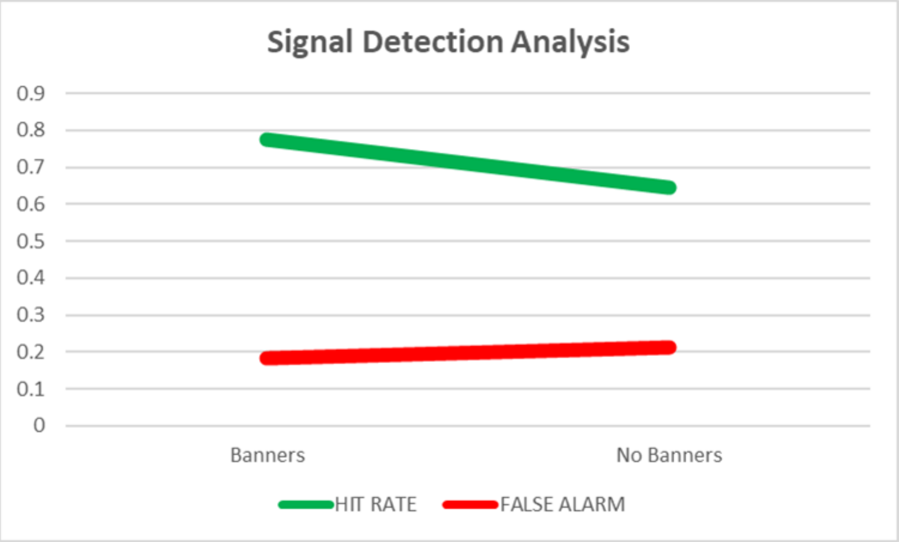

Figure 3: True and False Positive Rates - All Emails

ALL EMAILS	HIT RATE	FALSE ALARM	d'	λ
BANNERS	0.7754	0.1835	1.66	0.90
NO_BANNERS	0.6460	0.2135	1.17	0.79

RED	HIT RATE	FALSE ALARM	d'	λ
BANNERS	0.8025	0.1835	1.75	0.90
NO_BANNERS	0.5400	0.2135	0.89	0.79

The difference in d’ indicates that the Banner Group was better able to discriminate between safe and unsafe emails, when Red Banners were displayed, a difference of 0.86.

YELLOW	HIT RATE	FALSE ALARM	d'	λ
BANNERS	0.7688	0.1835	1.63	0.90
NO_BANNERS	0.6725	0.2135	1.24	0.79

The difference in d’ (0.39) between the Banners and No Banners groups when Yellow Banners were displayed was smaller than the Red Banners. The bias (λ) was the same for both Red and Yellow Banners. These differences in discriminability between the Red and Yellow Banners indicates that the Red Banners had a more significant impact on participants ability to identify potentially unsafe emails than the Yellow Banners. This is consistent with the perspective that Yellow Banner emails are more ambiguous in the threat they pose.

Summary

This study proved that INKY banners can significantly decrease risk within organizations. The banners provide just in time guidance to users to allow them to make a better decision as to the safety of a message they are reviewing. While this is intuitively obvious, this study provided empirical evidence of the potential return on investment provided by the INKY tools.

References

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. New York, NY: Routledge Academic.

About the author

Ira Winkler, CISSP, is President of Secure Mentem and Author of Advanced Persistent Security. He is considered one of the world's most influential security professionals and was named "The Awareness Crusader" by CSO magazine in receiving their CSO COMPASS Award. He has designed and implemented and supported security awareness programs at organizations of all sizes, in all industries, around the world.