Information security awareness training to combat social engineering attacks: A Meta-Analysis

Valerie Seinstra
The University of Twente
Enschede, The Netherlands
v.a.seinstra@student.utwente.nl

ABSTRACT

Information security awareness is crucial for equipping users to mitigate social engineering attacks effectively, where individuals are deceived into divulging sensitive data. Different training methods, such as games, lectures, or online textual training have been created to raise awareness and reduce vulnerability to these scams.

Building upon prior research conducted by Bullée and Junger (2020), this research performs a systematic literature review (N=22 effect sizes) and meta-analysis to find a more precise estimate of the true effect size of social engineering interventions between 2018 and 2023. Additionally, it compares the effects of different training methods and explores the potential influence of demographic factors.

The meta-analysis yielded an effect size of 0.161. However, the non-significant p-value of 0.141 at a significance level of 0.05 suggests the need for cautious interpretation.

KEYWORDS

Awareness training, Cybersecurity, Effectiveness, Information security awareness, Social engineering interventions

1 INTRODUCTION

Social engineering attacks use manipulation to persuade a victim to take a specific action to steal sensitive data [7]. Cybersecurity experts consider these attacks a major threat [28, 30]. Attackers use forms of deception and persuasion to exploit human vulnerabilities through social engineering and psychological techniques [28]. Findings from research [27] indicate that hackers can effectively and efficiently exploit victims' vulnerabilities, primarily due to the general lack of awareness of information security among the public. This makes the human link the most vulnerable [18, 27, 28]. However, interventions could lower this vulnerability and collectively help keep the organisation safe. Phishing is a type of social engineering attack [33]: if victims believe that a phishing link is genuine, they may unknowingly enter their login credentials into the fake website, giving attackers access to their account.

To mitigate the risk of falling for attacks, experts suggest raising awareness through training on identifying and avoiding them [18]. This is typically done in information security awareness training [1]. Information security awareness involves teaching users the necessary skills to protect themselves against social engineering attacks successfully [1]. Training companies offer services like simulated phishing attacks and educational materials to help customers combat social engineering threats [30]. Typically, the training is delivered through lecture and workshop-based methods, moderated by an expert with varying levels of involvement [11]. However, recently, serious games have been developed to provide this training too [18].

Generally, the effectiveness of this training is positive. However, most researchers utilized a small size of participants. Researchers [17] did a comparative literature review of studies between 2003 and 2018 on the effects of interventions. However, there have been more studies about this subject added to the body of knowledge. [24] conducted a similar study using a systematic literature review and PRISMA framework. They focused more on summarizing all articles instead of finding the effect size.

Bullée and Junger [7] researched the effectiveness of interventions based on studies until 2017 through a meta-analysis. This research follows up on their study. This would give a better insight into the effectiveness of current interventions. Furthermore, it compares different training methods and demographic factors on their influence. This research uses a systematic review and meta-analysis approach.

2 PROBLEM STATEMENT

There has been a lack of empirical evidence on the true effect size of social engineering interventions in the last five years, due to low participation sizes in most studies and the presence of contradictory findings. For example, research [10] concluded that the training was highly effective in reducing the number of users who fall victim to phishing attacks. This is in contradiction with research [3], which concluded that the training-receiving group was more likely to click on a fake link or submit personal data.

2.1 Research questions

This research tries to find an answer to the main research question:

- What is the effectiveness of information security awareness training in teaching people to recognize and combat social engineering attacks?

To answer the above question and examine the contributing factors, it will answer three sub-questions:

- How do different types of interventions differ in their effectiveness in reducing social engineering attacks?
- What is the impact of intervention characteristics on the effectiveness of information security awareness training?
- What is the impact of demographic factors on the effectiveness of information security awareness training?

39th Twente Student Conference on IT, July 07, 2023, Enschede, NL

https://doi.org/XXXXXXX.XXXXXXX

Based on these sub-questions, four coding categories have been created: context, characteristics of the intervention, methodology of the study, and demographics of the sample group. The first three categories are adopted from Bullée and Junger's study [7].

The *context* category describes the type of social engineering on which the study is based and which participants receive the training.

The *characteristics of the intervention* category is used to differentiate between the different characteristics of each training. This category is used to see if a certain training characteristic results in a higher effect size and, therefore, better training.

The *methodology of the study* category describes the manner in which the study is conducted. For example, the environment of the mock attack and the awareness of the participants could have a high impact on the effect size. In order to fairly compare the effect sizes of the studies, this information must be coded.

Lastly, the *demographics of the sample size* category is used to compare the age, gender and job position of the different sample groups.

3 RELATED WORK

Different studies have been conducted on the effect of social engineering interventions using different training methods. The methods used can be classified as games [5, 11, 13, 18, 33, 34], online training [3, 10, 11, 29, 30] and in-class lectures [10, 11, 20].

Types of interventions.
Some research has already been performed on the effectiveness of interventions. Researchers [3] conducted a study at a public research university located in the United States. Participants were grouped into two groups, with one group receiving an online-based "Cybersecurity Awareness Training". The effectiveness of the awareness program was evaluated using a quasi-experimental research design. They concluded that the training had a moderate effect in the opposite direction than was hypothesized. The training-receiving group was more likely to click on a fake link or submit personal data. According to the analysis, the older generation was found to be less susceptible to cyber deception than the younger generation. Additionally, the study suggested that people in higher job positions and those who have been employed for a longer period of time are more likely to fall for phishing attacks.

Other researchers [33] made a role-playing game called *What.Hack*. In this game, the players take the role of a bank employee [33]. The goal is to help the bank acquire contracts through emails without getting phished. Therefore, the players need to evaluate different emails within a limited time frame [33]. To evaluate the effectiveness of the game, the researchers analysed the correctness percentage, false negative rate, and false positive rate both before and after the game. They later compared the scores of the What.Hack game to two other role-playing games for anti-phishing training and found that the participants became more confident in their judgements after playing the What.Hack game.

Intervention characteristics.
Research [11] compared different training methods. It suggested that feedback is an essential component of an effective intervention. However, the outcomes of training on the participant's ability to recognize and mitigate threats vary depending on the delivery methods used.
Study [35] researched the effect of training using a role-playing scenario. The participants needed to identify different emails in the mailbox of Zhang Wei. Feedback was given to the participants after they identified each email to help them see the differences between phishing and legitimate emails. They found that the hit rate increased after adding the feedback and concluded that using a role-playing scenario with feedback improved the email identification results.

Demographics.
Study [19] researched the difference between employees with high and low job positions, for example, general employees vs managers. They provided their training using a lecture. They found that after training, the low-position group performed better than the control group, whereas the high-position group performed similarly regardless of the training condition.
Study [20] researched the difference in effectiveness between sex and age. They gave a lecture at primary schools and found that there was no significant effect of sex on the effect size. However, they found that there was a significant effect on age: the older pupils scored higher than the younger ones. Lastly, they concluded that the training had a medium effect size [21].

These studies illuminate crucial intervention techniques, and this research utilizes these findings to assess and identify the methods and characteristics that yield the highest effect size.

Research done by Bullée and Junger [7] concluded that interventions could be beneficial for lowering the chances of a successful attack. They based their work on research done before 2018. This research will add to their findings by analysing academic papers from 2018 onwards.

4 METHODOLOGY

This research uses a systematic review to collect and summarize all empirical evidence that fits the pre-specified eligibility criteria, previously used by Bullée and Junger [7]. Then, a meta-analysis is performed to summarize the results of these studies.

4.1 Data Collection

Scopus is used as database in the systematic literature review. This database is widely used by researchers and scholars, has comprehensive coverage and offers high-quality content [31]. The database was queried on May 4th 2023, with the following query:

KEY (("social engineering") OR (phishing) OR ((disclosure) AND ((cybercrime) OR (prevention))) AND ((experiment) OR (training) OR (survey) OR (warning) OR (intervention))) AND PUBYEAR > 2017 AND NOT TITLE-ABS-KEY ("neural network*" OR "deep learning" OR "machine learning") AND (EXCLUDE (SUBJAREA, "MEDI")) AND (LIMIT-TO (PUBSTAGE, "final")) AND (LIMIT TO (DOCTYPE, "cp") OR LIMIT TO (DOCTYPE, "ar") OR LIMIT TO (DOCTYPE, "re")) AND (LIMIT TO (LANGUAGE, "English") AND (LANGUAGE, "Dutch")).*

This query is based on the query employed in the study conducted by Bullée and Junger [7], with the exception that three

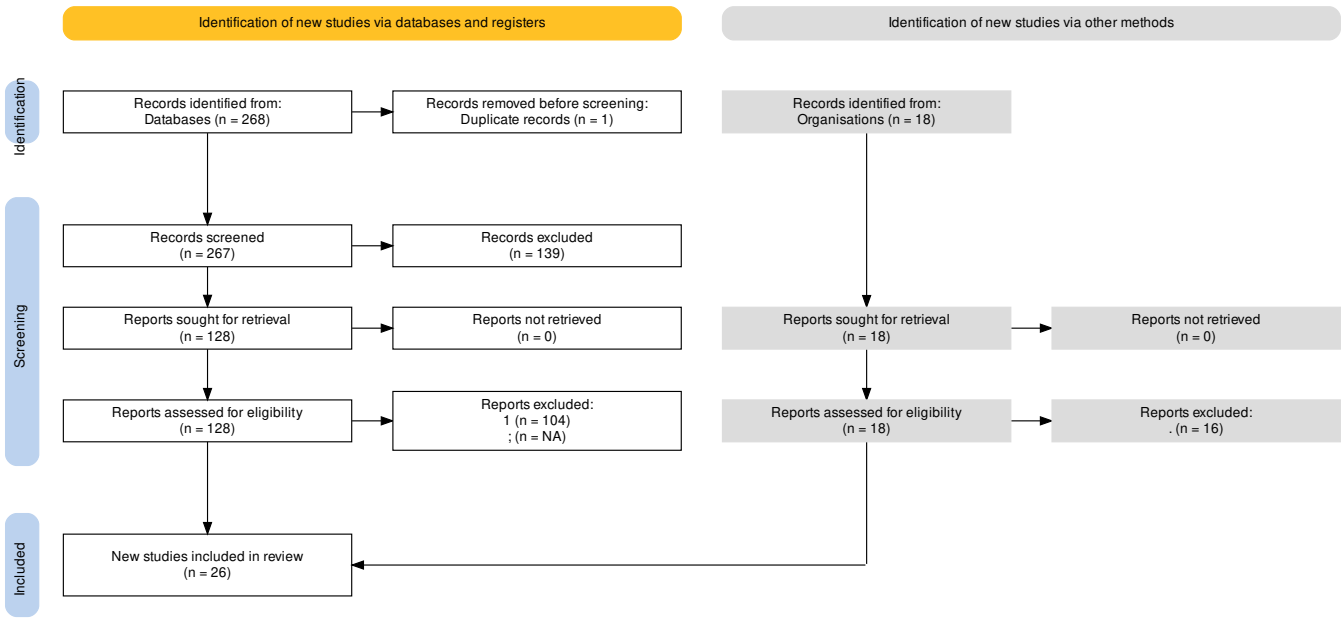


Figure 1: PRISMA Flow diagram [14]

eligibility criteria are included, namely, criteria 1, 2 and 9. Furthermore, the researcher noticed that many articles were about neural networks and machine learning. Therefore, these words were excluded from the search.

This research adopts Bullée and Junger's [7] eligibility criteria, except for criterion 9, which is modified to only include studies published after 2017.

Elegibility Criteria:

(1) To be a published scientific paper or a PhD thesis;
(2) The manuscript must be written in English or Dutch;
(3) The study should involve human subjects;
(4) An experimental design should be used, questionnaires or surveys that only measure, e.g. attitude or intention are excluded; it is of particular interest to observe how the subjects behave in the context of social engineering;
(5) The experiment (and intervention) should aim to reduce victimisation by social engineering; there should be deception or a malicious part be involved;
(6) There should be a comparison of at least two groups, i.e.: a control and training or awareness group; or a pre-training and post-training group; the comparison of groups is required to state the effectiveness of an intervention;
(7) No technical solutions (e.g. an algorithm that filters possible phishing emails); this analysis is about human behaviour in social engineering; therefore, exclusively technical solutions are excluded;
(8) There should be at least 20 observations per group; this was chosen to have sufficient strength in the analysis and reduce the possibility of the observations based on random chance;
(9) The publication date is after 2017.

The search query returned 268 results, including one duplicate. A secondary search with studies from colleagues within the university resulted in 18 more articles, resulting in a total of 285 unique articles. First, the title and abstract of all 285 articles were screened for eligibility. 139 articles were excluded. Then, the remaining 146 articles were screened in full text. This phase excluded 120 articles. The remaining 26 articles were included in this study. Most articles (58, 48.7%) failed on criteria 5. Furthermore, 14 (11.6%) articles failed on criteria 4, 23 (19%) articles failed on criteria 6, 19 (15.6%) articles failed on criteria 7 and 6 (5%) articles failed on criteria 8. Figure 1 shows the PRISMA Flow Chart of this systematic review.

Unfortunately, not all articles had the necessary information for a meta-analysis. Therefore, 10 emails were sent to the authors requesting additional information. One author responded and this resulted in two extra effect sizes. Furthermore, 5 studies did not mention the form of training provided. These studies were not seen as valid due to this important lack of information and were left out of the analysis. All other 11 articles were coded using the coding variables described below and 22 effect sizes were found.

4.2 Data Analysis

To analyse the data, a meta-analysis is performed. Meta-analysis combines the results of multiple studies, calculates an overall effect size, and uses subgroup analysis to explore potential sources of heterogeneity or variation in the effectiveness of the training programs. A random-effects model was used in the analysis [6] because it recognizes that the studies in the analysis are drawn from a larger population of potential studies and that the true effect sizes may differ across these studies. The meta-analysis was performed using the IBM SPSS program.

4.3 Coding variables

Bullée and Junger [7] coded their independent variables using three broad categories. This research uses the same categories, while also including some demographic variables. The independent variables were coded into the following categories:

4.3.1 Context.

Training type. Trainings can be given in various format types. Based on the screened articles, three types were identified: Rule-based [2, 4, 8, 16, 19, 20, 22, 23], mindfulness [16, 23] and game [25, 28, 32]. Rule-based trainings can be lectures or texts focusing on best practices for avoiding social engineering attacks. Mindfulness training emphasizes pausing to consider the context of requests and engaging in active questioning when evaluating emails to identify suspicious elements [23]. It also encourages individuals to seek advice from trusted sources and gather evidence before making decisions about potentially suspicious emails [23]. Lastly, games incorporate interactive scenarios and challenges to educate participants about identifying and responding to phishing attempts effectively [28].

Type of social engineering. Social engineering attacks can reach their victims using different methods. Most studies developed mock attacks to evaluate the effect of their training. All coded studies used email phishing.

Pre-victimisation. Some studies only provided training when participants fell for their initial mock attack. This type of training is called embedded training. It only provides training to those who need it and motivates them to acquire the skills necessary for defending themselves against real attacks [7].

4.3.2 Characteristics of the intervention.

Modality intervention. Training can be delivered through various methods, such as oral presentations, static content like PDF documents, or dynamic approaches like games [7].

Priming. Priming refers to the technique used by attackers to create a context or mindset that makes their targets more susceptible to falling for the social engineering attack [12]. Priming techniques are used to manipulate the target's cognitive biases, emotions, and trust [12].

Warning. Warnings can be shown on a website or email to warn a user of harm. This could make them behave more safely [7]. In an educational context, warnings could make a user aware of the dangers of a website or email.

Focus. There are many different types of social engineering techniques. Therefore, most training programs focus on a specific technique or type. For example, training could specifically focus on email or URL characteristics [4, 8, 16, 26, 28, 32], or training programs could provide more general information about social engineering or cybercrime [2, 19, 23, 25].

Technical measures. Some training programs include extra security measures by incorporating additional layers of technical safeguards [7]. This makes participants unable to take certain actions, for example, open certain emails or visit certain websites.

Format. Programs could adapt different format methods to deliver their trainings. For instance, text messages, comics or games could be used [7]. The text could have graphics or not. Different program formats could differ in their effectiveness in reducing social engineering attacks.

Tips. Tips could be provided to combat social engineering attacks. Tips could be in the form of 'never click on links within emails' or 'find and call the real customer service' [7]. These tips focus on providing specific guidance to participants when they find themselves unsure about what actions to take in a given situation and could be considered best practices.

Intensity. Some interventions are more intense than others. Bullée and Junger [7] considered the intensity low when only information with tips is provided, medium when additional reading materials are given, and the intensity high when a lecture or game is included in the training.

4.3.3 Methodological aspects of the study.

Environment. The environment of the mock attack influences the heightened awareness of the participants that they are being tested [7]. It can be expected that participants that are tested in a lab environment perform better than participants that are being tested in a real-life situation [9].

Awareness. Participants can be aware or unaware that they are being tested [7]. As stated above, in a controlled lab environment, participants are expected to outperform those tested in a real-life setting, since their awareness of phishing is heightened [9].

Randomisation. To evaluate the validity of the study, randomisation is coded. The design of the study is important in evaluating the validity of the outcome [7]. Studies with a stronger internal validity tend to report weaker effectiveness effects than studies with a weaker internal validity [7].

4.3.4 Demographics of the sample group.

Age. Several studies have examined the influence of age on different training formats, and it is anticipated that younger age groups would likely learn more from games, whereas older age groups would be expected to benefit more from lectures [20].

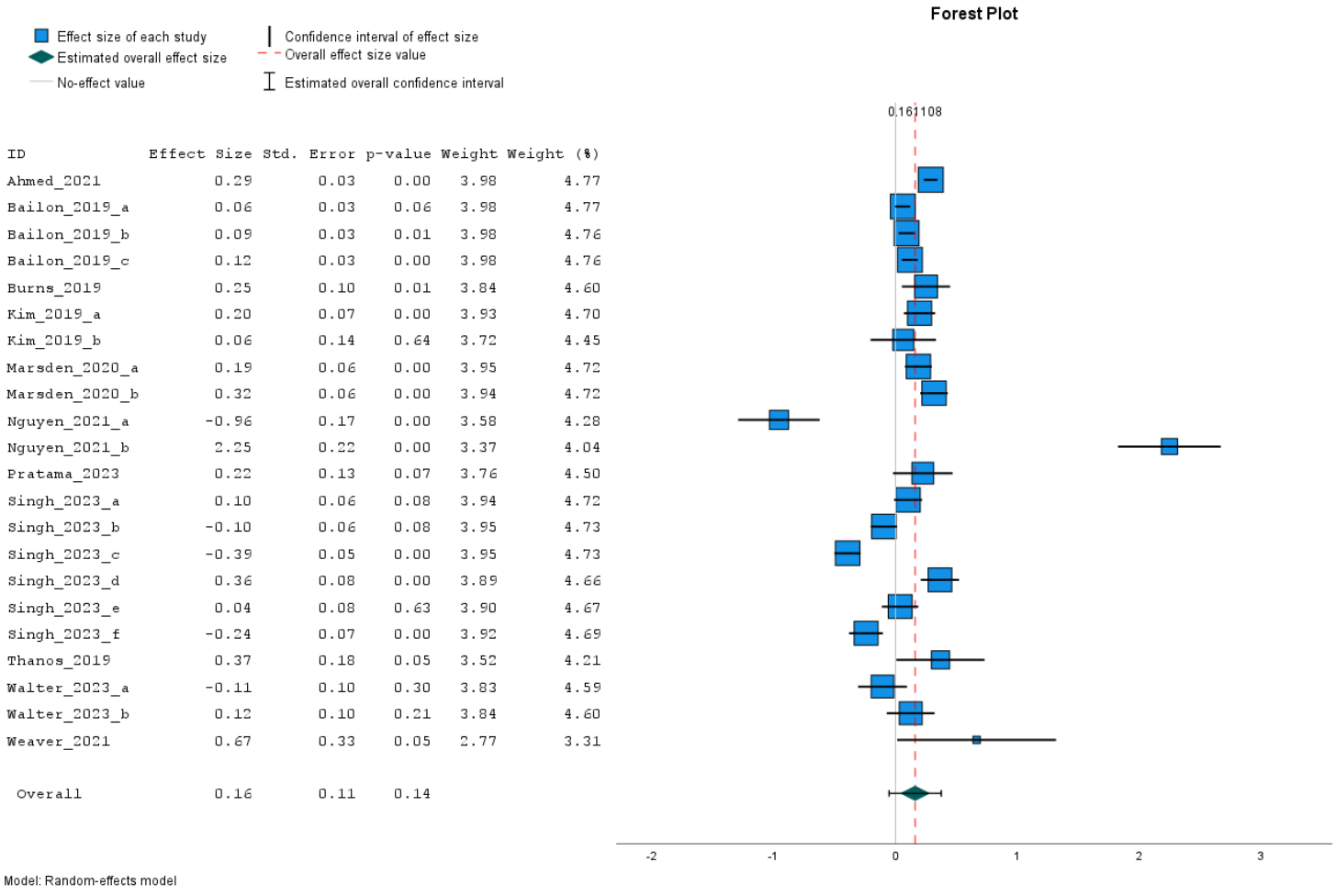
Gender. Gender could play a role in the effectiveness of different training formats. Some studies compared the training outcomes based on gender, as is done in research [26].

Job position. Different employee layers in a company could have different effectiveness outcomes on the mock attack [19]. This research tries to compare three job positions: students, employees and managers.

5 RESULTS

5.1 Overall result

The meta-analysis included 11 studies [2, 4, 8, 16, 19, 22, 23, 25, 26, 28, 32], having 22 effect sizes (see Figure 2). Cohen's d was used as the effect size measurement [21]. The overall effect size is 0.161 (95% CI -0.053, 0.376) with a standard error of 0.1094. According to



ID	Effect Size	Std. Error	p-value	Weight	Weight (%)
Ahmed_2021	0.29	0.03	0.00	3.98	4.77
Bailon_2019_a	0.06	0.03	0.06	3.98	4.77
Bailon_2019_b	0.09	0.03	0.01	3.98	4.76
Bailon_2019_c	0.12	0.03	0.00	3.98	4.76
Burns_2019	0.25	0.10	0.01	3.84	4.60
Kim_2019_a	0.20	0.07	0.00	3.93	4.70
Kim_2019_b	0.06	0.14	0.64	3.72	4.45
Marsden_2020_a	0.19	0.06	0.00	3.95	4.72
Marsden_2020_b	0.32	0.06	0.00	3.94	4.72
Nguyen_2021_a	-0.96	0.17	0.00	3.58	4.28
Nguyen_2021_b	2.25	0.22	0.00	3.37	4.04
Pratama_2023	0.22	0.13	0.07	3.76	4.50
Singh_2023_a	0.10	0.06	0.08	3.94	4.72
Singh_2023_b	-0.10	0.06	0.08	3.95	4.73
Singh_2023_c	-0.39	0.05	0.00	3.95	4.73
Singh_2023_d	0.36	0.08	0.00	3.89	4.66
Singh_2023_e	0.04	0.08	0.63	3.90	4.67
Singh_2023_f	-0.24	0.07	0.00	3.92	4.69
Thanos_2019	0.37	0.18	0.05	3.52	4.21
Walter_2023_a	-0.11	0.10	0.30	3.83	4.59
Walter_2023_b	0.12	0.10	0.21	3.84	4.60
Weaver_2021	0.67	0.33	0.05	2.77	3.31
Overall	0.16	0.11	0.14		

Model: Random-effects model

Figure 2: Forest Plot

Cohen, this effect size can be considered trivial [21]. The Z-score for this meta-analysis is 1.472 and the p-value is 0.141.

5.2 Subgroup analysis - Context

The results of the subgroup analysis can be found in Appendix A.

Training type. The types of training differ greatly in effectiveness. Rule-based training and game have an effect size of 0.102 and 0.046 respectively. These effects can be considered trivial [21]. Mindfulness has a large effect (1.064).

Type of social engineering. Since there was only one type of social engineering found in the papers, namely email, this research could not do a subgroup analysis on the type of social engineering variable.

Pre-victimised. The effect sizes of pre-victimised and not pre-victimised were 0.286 and 0.152 respectively. The former can be considered small and the latter can be considered trivial [21].

5.3 Subgroup analysis - Characteristics of the intervention

Modality. There were three different types of modality found in the studies: orally, static content and dynamic. Orally has a small effect (0.304), static content has an effect that can be considered trivial (0.133), and dynamic has a trivial effect size (0.046) [21].

Priming. Training programs that included priming in their mock attack had an effect size of 0.093, compared to no priming with an effect size of 0.252. This means that the effect of priming can be considered trivial and no priming a small effect [21].

Warning. All studies did not have any warnings in their interventions. Therefore, this research could not do a subgroup analysis on the warning variable.

Focus. Most studies focused on both URL and email, or social engineering in general. 'Both URL and email' and other focuses have effect sizes that can be considered trivial (0.042 and 0.104 respectively) [21], while social engineering has a small effect size (0.362).

Technical. All studies did not have any technical help in their interventions. Therefore, this research could not do a subgroup analysis on the technical variable.

Format. The effect size of trainings that included text was 0.088, which can be considered trivial [21]. Furthermore, text plus graphics had an effect size of 0.298, which can be considered small [21]. The effect size of comic suggests that there was no improvement but a worse performance after the training (-0.107). Lastly, the effect size of trainings that used a game was 0.046, which can be considered a trivial effect [21].

Tips. Trainings that used no tips had a negative effect size (-0.730), which indicates no improvement and worse performance. Tips with additional materials had an effect size that can be considered trivial (0.145) [21]. Furthermore, trainings that used only tips had a small effect size (0.262).

Intensity. Trainings with a low intensity had a small effect (0.282), while a medium and intense intensity only had a trivial effect (0.103 and 0.093 respectively) [21].

5.4 Subgroup analysis - Methodological aspects

Environment. Trainings with a mock attack in a lab environment had an effect size of 0.038, while a mock attack conducted in the wild had an effect size of 0.288. This means the lab had a trivial effect and the wild had a small effect [21].

Awareness. The mock attacks where the participants were aware that they were being tested provided an effect size of 0.161, which can be considered trivial [21]. Where participants were unaware resulted in a trivial effect (0.180).

Random. Studies that randomly assigned their participants had a lower effect size than the studies that did not randomly assign their participants (0.166 and 0.172 respectively).

5.5 Subgroup analysis - Demographics

Age. There were no results found for age in the studies. No studies compared the age of their participants to the outcome. Therefore, this research could not do a subgroup analysis on age.

Gender. There were no results found for gender in the studies. No studies compared the gender of their participants to the outcome. Therefore, this research could not do a subgroup analysis on gender.

Job position. The effect sizes of the four job positions were as follows: student (0.484, medium effect), employee (0.166, trivial effect), manager (0.064, trivial effect), and unknown position (-0.006, no effect).

6 CONCLUSION

The effectiveness of information security awareness training in teaching people to recognize and combat social engineering attacks is trivial considering an effect size of 0.161. This means that the observed difference between the variables in this study is very small or negligible, and the statistical analysis did not find enough evidence to support the presence of a significant effect.

Cochran's Q statistic was used to test for homogeneity in the subgroup analysis. The Q statistic and its associated p-value provide information about the homogeneity of effect sizes across subgroups. A significant p-value indicates heterogeneity, while a non-significant p-value suggests homogeneity among the subgroups. This means that, if the p-value is significant, the subgroups differ from each other and one may be more effective than the others.

6.1 RQ1: Intervention type

Training types' Q statistic of the test of subgroup homogeneity is not significant on a 0.05 level. This indicates that there is no significant heterogeneity among the subgroups, supporting the assumption of homogeneity. This means that the effect sizes are relatively consistent across the subgroups, and any observed differences can be attributed to chance.

The Q statistic's significant level of pre-victimisation is also not significant. Therefore, this too supports the assumption of homogeneity. This means that there is not a significant difference between the subgroups.

6.2 RQ2: Intervention characteristics

Almost all intervention characteristics do not have a significant Q statistic, except for format. These non-significant Q statistics support the assumption of homogeneity and there is not a significant difference between the subgroups.

However, format does have a significant Q statistic at a 0.05 level. It suggests significant heterogeneity among the subgroups. This indicates that the effect sizes are not consistent across the subgroups. Text plus graphics has the highest effect size, although this can be considered small [21]. This means that this format is superior to the other subgroups, namely only text, comic, or game. According to this analysis, for the best results, the design of an intervention should have a format that includes text and graphics.

6.3 RQ3: Demographics

Since there were no statistics about age and gender, there can be no conclusion drawn. However, there can be conclusions made about job position. Students had the highest effect size (medium) [21]. However, the Q statistic is not significant, suggesting homogeneity. This means that there is no significant difference between the subgroups.

7 DISCUSSION

The study conducted by Bullée and Junger [7] provided an effect size of 0.54. This research calculated a significantly lower effect size of 0.161. Factors such as advancements in technology, evolving social engineering tactics, and increased awareness among individuals may have influenced the effectiveness of training programs. Therefore, the effectiveness observed in the study of Bullée and Junger may not directly translate to the more recent timeframe covered in this study. Furthermore, the baseline knowledge, attitudes, or behaviours related to information security of the participants might be changed. In recent years, there has been a notable increase in campaigns and extensive media coverage, resulting in heightened public awareness regarding the risks and consequences involved [15]. These differences can influence the effectiveness of

the training since participants begin with higher awareness and this can result in varying effect sizes.

There can be two outliers noticed in the analysis. These resulted in a large overall confidence interval, which includes the null value. Both outliers were from the same study [23], which could indicate some biases or differences in that study.

After carefully analysing the subgroup analysis, some variables stand out.

Firstly, it is noteworthy that the intensity variable yielded opposite effect sizes than expected. It was expected that an intense training yielded a higher effect size than a lower intensity. However, the results show that the lowest intensity has the highest effect size. Furthermore, the insignificant Q statistic suggests that the intensity of training may not be a determining factor in the effectiveness of the intervention. This finding may warrant further investigation or exploration of other factors that could contribute to the observed effect sizes in the different training intensities.

Secondly, it was expected that mock attacks conducted in a lab resulted in a higher effect size. This research's results show that a mock attack in the wild yielded the highest effect size.

Looking at the random variable, the results of the analysis are intriguing as the two effect sizes obtained are remarkably similar. This is contrary to the expectation that studies with less stringent designs would show better outcomes compared to studies with stricter designs. Surprisingly, in this study, the effect sizes from both types of studies exhibited minimal differences.

Lastly, it was expected that participants that were aware that they were being tested had a higher effect size. However, this research resulted in the unaware participants having the highest effect size.

It can be recommended to Chief Information Security Officers to improve security for the organisation by designing an intervention that includes text and graphics, although the effectiveness of these interventions is very small or negligible based on the calculated results.

8 LIMITATIONS

The limitations of this research include time pressure and availability. Many more effect sizes could be found given the initial 26 studies. However, given the time pressure, the authors had only 3 weeks to respond to the emails. With more time, more authors might have responded. Additionally, no included studies have explored the differences based on age and gender. Consequently, no specific findings or results regarding these factors can be provided at this time. Lastly, there were many variables coded, resulting in many variables only having 1 or 2 studies that were applicable. These outcomes could be more precise with more studies added.

9 FUTURE RESEARCH

It would be valuable for future research to explore the reasons behind the calculated insignificant Q statistics across all variables in the subgroup analysis. Understanding why the effect sizes do not significantly differ based on various factors can provide deeper insights into the complex nature of social engineering interventions and their impact. It may reveal additional variables, moderators, or contextual factors that influence the effectiveness of the interventions, leading to a more comprehensive understanding of the training outcomes and the development of more effective training strategies in the future.

REFERENCES

[1] Abdul Rahman Ahlan, Muharman Lubis, and Arif Ridho Lubis. Information security awareness at the knowledge-based institution: Its antecedents and measures. *Procedia Computer Science*, 72:361–373, 2015. doi: 10.1016/j.procs.2015.12.151.

[2] V. Ahmed and S. Al-Haddad. The use of social engineering to change organizational behavior toward information security in an educational institution. *Journal of Information System Security*, 17(2):103–125, 2021.

[3] S. Back and R.T. Guerette. Cyber place management and crime prevention: The effectiveness of cybersecurity awareness training against phishing attacks. *Journal of Contemporary Criminal Justice*, 37:427–451, 2021. doi: 10.1177/10439862211001628.

[4] Aurélien Baillon, Jeroen de Bruin, Aysil Emirmahmutoglu, Evelien van de Veer, and Bram van Dijk. Informing, simulating experience, or both: A field experiment on phishing risks. *PLOS ONE*, 14(12), 2019. doi: 10.1371/journal.pone.0224216.

[5] Ahmed Baiomy, Mahmoud Mostafa, and Alyaa Youssif. Anti-phishing game framework to educate arabic users: Avoidance of urls phishing attacks. *Indian Journal of Science and Technology*, 12:01–10, 11 2019. ISSN 09746846. doi: 10.17485/ijst/2019/v12i44/147850.

[6] Michael Borenstein, Larry V. Hedges, Julian P.T. Higgins, and Hannah R. Rothstein. A basic introduction to fixed-effect and random-effects models for tell me-analysis. *Research Synthesis Methods*, 1(2):97–111, 2010. doi: 10.1002/jrsm.12.

[7] Jan Willem Bullee and Marianne Junger. How effective are social engineering interventions? a tell me-analysis. *Information and Computer Security*, 28:801–830, 11 2020. ISSN 2056497X. doi: 10.1108/ICS-07-2019-0078.

[8] A. J. Burns, M. Eric Johnson, and Deanna D. Caputo. Spear phishing in a barrel: Insights from a targeted phishing campaign. *Journal of Organizational Computing and Electronic Commerce*, 29(1):24–39, 2019. doi: 10.1080/10919392.2019.1552745.

[9] Rebecca M. Calisi and George E. Bentley. Lab and field experiments: Are they the same animal? *Hormones and Behavior*, 56(1):1–10, 2009. doi: 10.1016/j.yhbeh.2009.02.010.

[10] Anthony Carella, Murat Kotsoev, and Traian Marius Truta. Impact of security awareness training on phishing click-through rates. *2017 IEEE International Conference on Big Data (Big Data)*, 2017. doi: 10.1109/bigdata.2017.8258485.

[11] A. Darem. Anti-phishing awareness delivery methods. *Engineering, Technology, Applied Science Research*, 11(6):7944–7949, 2021. doi: 10.48084/etasr.4600.

[12] Edwin D. Frauenstein and Stephen V. Flowerday. Social network phishing: Becoming habituated to clicks and ignorant to threats? *2016 Information Security for South Africa (ISSA)*, 2016. doi: 10.1109/issa.2016.7802935.

[13] Matthew J. Grubbs. Anti-phishing game-based training: An experimental analysis of demographic factors. *SSRN Electronic Journal*, 2022. doi: 10.2139/ssrn.4011558.

[14] Neal R. Haddaway, Matthew J. Page, Chris C. Pritchard, and Luke A. McGuinness. Prisma2020: An r package and shiny app for producing prisma 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Systematic Reviews*, 18(2), 2022. doi: 10.1002/cl2.1230.

[15] Julian Hazell. Large language models can be used to effectively scale spear phishing campaigns, May 2023. URL https://doi.org/10.48550/arXiv.2305.06972.

[16] David Michael Hull, Sebastian Walter Schuetz, and Paul Benjamin Lowry. Tell me a story: The effects that narratives exert on meaningful-engagement outcomes in antiphishing training. *Computers Security*, 129:103252, 2023. doi: 10.1016/j.cose.2023.103252.

[17] Daniel Jampen, Gürkan Gür, Thomas Sutter, and Bernhard Tellenbach. Don't click: Towards an effective anti-phishing training. a comparative literature review. *Human-centric Computing and Information Sciences*, 10(1), 2020. doi: 10.1186/s13673-020-00237-7.

[18] Gokul Jayakrishnan, Vijayanand Banahatti, and Sachin Lodha. Pickmail: A serious game for email phishing awareness training. *Usable Security and Privacy Symposium 2022*, 2022. doi: 10.14722/usec.2022.23059.

[19] Bora Kim, Do-Yeon Lee, and Beomsoo Kim. Deterrent effects of punishment and training on insider security threats: A field experiment on phishing attacks. *Behaviour Information Technology*, 39(11):1156–1175, 2019. doi: 10.1080/0144929x.2019.1653992.

[20] E. Lastdrager, I.C. Gallardo, P. Hartel, and M. Junger. How effective is anti-phishing training for children? *Proceedings of the 13th Symposium on Usable Privacy and Security, SOUPS 2017*, pages 229–239, 2019.

[21] C. W. LeCroy and J. Krysik. Understanding and interpreting effect size measures. *Social Work Research*, 31(4):243–248, 2007. doi: 10.1093/swr/31.4.243.

[22] John Marsden, Zachary Albrecht, Paula Berggren, Jessica Halbert, Kyle Lemons, Anthony Moncivais, and Matthew Thompson. Facts and stories in phishing training: A replication and extension. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020. doi: 10.1145/3334480.3381435.

[23] Christopher Nguyen, Matthew Jensen, and Eric Day. Learning not to take the bait: A longitudinal examination of digital training methods and overlearning on phishing susceptibility. *European Journal of Information Systems*, 32(2):238–262,

2021. doi: 10.1080/0960085x.2021.1931494.

[24] Kanchan Patil and Sai Rohith Arra. Detection of phishing and user awareness training in information security: A systematic literature review. *Proceedings of 2nd International Conference on Innovative Practices in Technology and Management, ICIPTM 2022*, pages 780–786, 2022. doi: 10.1109/ICIPTM54933.2022.9753912.

[25] Georgios Pouraimis, Konstantinos Thanos, Athanasios Grigoriadis, and Stelios C. Thomopoulos. Long lasting effects of awareness training methods on reducing overall cyber security risk. *Signal Processing, Sensor/Information Fusion, and Target Recognition XXVIII*, 2019. doi: 10.1117/12.2518934.

[26] Ahmad R. Pratama, Nunu Vadila, and Firman M. Firmansyah. Exposing generational and gender gap in phishing awareness among young adults: A survey experiment. *VII INTERNATIONAL CONFERENCE "SAFETY PROBLEMS OF CIVIL ENGINEERING CRITICAL INFRASTRUCTURES" (SPCECI2021)*, 2023. doi: 10.1063/5.0114868.

[27] Tanusree Sharma and Masooda Bashir. An analysis of phishing emails and how the human vulnerabilities are exploited. *Advances in Intelligent Systems and Computing*, 1219 AISC:49–55, 2020. ISSN 21945365. doi: 10.1007/978-3-030-52581-1_7.

[28] 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. doi: 10.1177/1071181320641097.

[29] A. Sumner, X. Yuan, M. Anwar, and M. McBride. Examining factors impacting the effectiveness of anti-phishing trainings. *Journal of Computer Information Systems*, 62:975–997, 2022. doi: 10.1080/08874417.2021.1955638.

[30] Thomas Sutter, Ahmet Selman Bozkir, Benjamin Gehring, and Peter Berlich. Avoiding the hook: Influential factors of phishing awareness training on click-rates and a data-driven approach to predict email difficulty perception. *IEEE Access*, 10:100540–100565, 2022. ISSN 21693536. doi: 10.1109/ACCESS.2022.3207272.

[31] Elizabeth S. Vieira and José A. Gomes. A comparison of scopus and web of science for a typical university. *Scientometrics*, 81(2):587–600, 2009. doi: 10.1007/s11192-009-2178-0.

[32] Bradley W. Weaver, Adam M. Braly, and David M. Lane. Training users to identify phishing emails. *Journal of Educational Computing Research*, 59(6):1169–1183, 2021. doi: 10.1177/0735633121992516.

[33] Z.A. Wen, Z. Lin, R. Chen, and E. Andersen. What.hack: Engaging anti-phishing training through a role-playing phishing simulation game. *Conference on Human Factors in Computing Systems - Proceedings*, 2019. doi: 10.1145/3290605.3300338.

[34] Dikka Aditya Wibawa, Hermawan Setiawan, and Girinoto. Anti-phishing game framework based on extended design play experience (dpe) framework as an educational media. *2022 7th International Workshop on Big Data and Information Security (IWBIS)*, 2022. doi: 10.1109/iwbis56557.2022.9924935.

[35] Ying Zhou, Xinyue Cui, Weina Qu, and Yan Ge. The effect of automation trust tendency, system reliability and feedback on users' phishing detection. *Applied Ergonomics*, 102:103754, 2022. doi: 10.1016/j.apergo.2022.103754.

A APPENDIX

Context							
Training_type	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
Game	0.046	0.1094	0.416	0.677	-0.169	0.260	8
Mindfulness	1.064	1.1785	0.903	0.367	-1.246	3.374	2
Rule-based	0.102	0.0789	1.292	0.197	-0.053	0.256	12
Homogeneity	Q = 0.866		Sig. = 0.649				
Type of SE	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
Email	0.161	0.1094	1.472	0.141	-0.053	0.376	22
Pre-victimisation	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
No	0.152	0.1218	1.244	0.213	-0.087	0.390	20
Yes	0.286	0.0272	10.507	0.000	0.233	0.340	2
Homogeneity	Q = 1.163		Sig. = 0.281				
Characteristics of the intervention							
Modality	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
Orally	0.304	0.3518	0.865	0.387	-0.285	0.994	7
Static content	0.133	0.0465	2.862	0.004	0.042	0.224	7
Dynamic	0.046	0.1094	0.416	0.677	-0.169	0.260	8
Homogeneity	Q = 0.813		Sig. = 0.666				
Priming	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
No	0.252	0.3074	0.819	0.413	-0.351	0.854	8
Yes	0.093	0.0628	1.474	0.140	-0.030	0.216	14
Homogeneity	Q = 0.257		Sig. = 0.612				
Warning	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
No	0.161	0.1094	1.472	0.141	-0.053	0.376	22
Focus	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
Both URL and email	0.042	0.0642	0.602	0.547	-0.094	0.178	12
Social engineering	0.362	0.4182	0.865	0.387	-0.458	1.182	6
Other	0.104	0.0235	4.424	<,0.001	0.058	0.150	2
Homogeneity	Q = 6.208		Sig. = 0.102				
Technical	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
No	0.161	0.1094	1.472	0.141	-0.053	0.376	22
Format	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
Tekst	0.088	0.0188	4.672	<,0.001	0.051	0.125	3
Tekst + graphics	0.298	0.3033	0.982	0.326	-0.297	0.892	8
Comic	-0.107	0.1025	-1.044	0.296	-0.308	0.094	1
Game, including quiz or Q&A	0.046	0.1094	0.416	0.677	-0.169	0.230	8
Homogeneity	Q = 10.608		Sig. = 0.031				
Tips	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
No	-0.730	0.1156	-0.636	0.525	-0.300	0.153	4
Only tips	0.262	0.2726	0.960	0.337	-0.273	0.796	9
Tips with additional materials	0.145	0.0458	3.168	0.002	0.055	0.235	9
Homogeneity	Q = 3.369		Sig. = 0.186				
Intensity	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
Low	0.282	0.3524	0.800	0.423	-0.409	0.973	13
Medium	0.103	0.1976	0.522	0.601	-0.284	0.491	2
Intense	0.093	0.0696	1.343	0.179	-0.043	0.230	7
Homogeneity	Q = 0.276		Sig. = 0.871				
Methodological aspects							
Environment	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
Lab	0.038	0.6970	0.552	0.581	-0.098	0.175	12
Wild	0.288	0.2365	1.217	0.224	-0.176	0.751	10
Homogeneity	Q = 1.023		Sig. = 0.312				
Awareness	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
No	0.180	0.0340	5.301	<,0.001	0.113	0.247	10
Yes	0.161	0.2144	0.691	0.489	-0.272	0.569	12
Homogeneity	Q = 0.021		Sig. = 0.884				
Random	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
No	0.172	0.0592	2.912	0.004	0.056	0.288	2
Yes	0.166	0.1219	1.360	0.141	-0.073	0.405	22
Homogeneity	Q = 0.002		Sig. = 0.962				
Demographics							
Job position	**Effect size**	**Std. Error**	**Z**	**Sig. (2-tailed)**	**Lower**	**Upper**	**N**
Student	0.484	0.5165	0.936	0.349	-0.529	1.496	5
Employee	0.166	0.0399	4.163	<,0.001	0.088	0.245	7
Manager	0.064	0.1371	0.465	0.642	-0.205	0.332	1
Unknown	-0.006	0.0798	-0.072	0.942	-0.162	0.151	9
Homogeneity	Q = 4.419		Sig. = 0.220				