Effectiveness of Social Engineering Interventions over time

Robbert Derksen
University of Twente
PO Box 217, 7500 AE Enschede
the Netherlands

r.derksen@student.utwente.nl

ABSTRACT

With the increase of cybercrime every year and the costs to society associated with it, there is a need to minimise the effectiveness of these crimes. Developing better security software is one side of the coin, but cybercriminals often use social engineering attacks to accomplish their goals. These attacks lure the computer user into granting access and compromising the system. Researchers have developed interventions to weapon people against social engineering attacks. This research investigates the effectiveness of various aspects of current interventions. We do so by conducting a meta-analysis and a subgroup analysis on the existing literature. Ultimately, a regression analysis is done to find the effect of time on the interventions. All studies on Scopus from 2015 till 2021 related to social engineering were scanned for relevance. Studies needed to have an experimental design with human subjects, focussing on reducing victimisation using an intervention. We found that the current interventions have a medium effect size (SMD = 0.503), but some interventions are more effective than others. Interventions with the highest effect size had a dynamic modality, included tips and warnings in combination with training material and focussed on recognising fraudulent URL's. Interventions had a medium or high intensity. The regression analysis showed a decrease in effect size as time went by. However, the findings were not statistically significant enough to conclude that interventions lose their effectiveness over time.

Keywords

Cybercrime, Meta-analysis, Phishing, Phishing training, Social engineering, Social engineering interventions, Qualitative synthesis

1. INTRODUCTION

Over the past decade, the amount of computer usage has vastly increased. This contributed to a climate where cybercrime is more popular among criminals. According to the Anti-Phishing Working group research, the number of reported phishing attacks has doubled since early 2020 [1]. Phishing is a type of cybercrime that falls under social engineering. Social engineering is an attack vector wherein attackers try to exploit human weaknesses through social interaction to compromise a computer [23].

'Social engineering schemes prey on unwary victims by fooling them into believing they are dealing with a trusted, legitimate party, such as by using deceptive email addresses and email messages. These are designed to lead consumers to counterfeit Websites that trick recipients into divulging financial data such as usernames and passwords [1]. Social engineering is an effective cyberattack because it targets the human behind the system. Humans are considered the weakest link in cybersecurity [4][16][18], and therefore, it seems sensible from the point of cybercriminals to use social engineering attacks instead of attacking the system directly.

The average cost of data breaches caused by social engineering to businesses is 4.47 million dollars [11], so it is in their best interest to prevent these breaches from occurring. Researchers have been developing interventions, and organisations use them to weapon their employees against social engineering attacks. Interventions can vary from intensive classroom training [17] to a few rules on paper to adhere to [9]. Some interventions appear to be effective in reducing the likelihood of being victimised by an attack [2][21], while others do not show a significant effect [14] [24]. Some interventions even show a negative effect [13]. Ultimately, some interventions seem to lose their effectiveness over time [3][6][15].

2. PROBLEM STATEMENT

With the increase in cybercrime and the costs associated with a data breach, it is in society's interest to develop interventions that significantly reduce the likelihood of being victimised by social engineering attacks. This study investigates the effectiveness of social engineering interventions and their retention. By doing so, we aim to provide a blueprint for future interventions.

2.1 Research questions

This study answers the following research question:
RQ1: What are good social engineering interventions, and how often should they be repeated?
The questions were broken down into three sub-questions to answer the research question.
SQ1: How effective are social engineering interventions?
SQ2: What characteristics of interventions are effective?
SQ3: What is the retention of social engineering interventions?

We can build a blueprint for future interventions by answering the three sub-questions.

3. RELATED WORK

Various studies used an experimental design with human subjects to measure the effectiveness of social engineering interventions. Bullée et al. (2016) [6] researched how susceptive people are to a technical support scam and the effect of time between intervention and compliance. They found that subjects receiving the intervention had eight times lower odds of being victimised. However, the effectiveness did not last for more than two weeks. Arachchilage et al. (2016) [2] measured the efficacy

36th Twente Student Conference on IT, Febr. 4 , 2022, Enschede, The Netherlands. Copyright 2021, University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science.

of a game prototype among computer science students. Students scored 28% better after the intervention. Jensen et al. (2017) [12] and Nguyen et al. (2021) [17] compared the effect of rule-based interventions versus interventions focused on mindfulness. In both studies, subjects receiving mindfulness training were less susceptible to phishing attacks. In addition, Jensen et al. (2017) [12] also researched the effectiveness of interventions consisting of text versus interventions composed of text and graphics. They did not find a significant difference between the two formats. Junger et al. (2017) [13] studied the effect of priming and warnings as aspects of social engineering interventions. Both priming and warnings did not have a positive effect. Lastdrager et al. (2017) [15] researched the effectiveness of interventions for children in a classroom setting and their knowledge retention. Childrens' ability to recognise phishing emails increased after the classroom training, but the effect lasted only four weeks. Stockhardt et al. (2016) [21] compared the effectiveness of text-based, computer-based, and instructor-based interventions. Instructor-based training received the best results but was not time efficient. Interestingly, text-based training outperformed computer-based training. All but one of the studies mentioned above were included in Bullée and Junger's meta-analysis (2020) [5]. They found medium effectiveness (SMD = 0,54) and knowledge retention (ß = -0,0005), indicating that the effect size decreases with 0,0005 for every hour between the intervention and the attack.

Since then, more research about the effectiveness of social engineering has been done. Yang et al. (2017) [26] researched the effect of warning interfaces next to a basic phishing training. They found that their training was only effective in combination with warning interfaces. Wash and Cooper (2018) [24] used a 2x2 design to measure the effect of advice versus stories and experts versus non-experts. 'facts-and advice led to lower click rates when appearing to come from an expert, but stories led to lower click rates when appearing to come from peers rather than experts'. Kim et al. (2019) [14] compared the effect of pre-victimisation and being punished against the effect of cybersecurity training. Receiving a punishment or training did decrease the likelihood of being victimised. Baillon et al. (2019) [3] did something similar and measured the effect of three different interventions. A training group received information about phishing and how one can recognise a phishing email. An 'experience' group received a phishing email and was briefed about it. And a third group received both the information and the experience. The groups that received information and a phishing email did both have a significant effect, but the group that received both interventions did not score better than the individual groups. Tschakert and Ngamsuriyaroj (2019) [22] researched the effectiveness of a combination of text-based, video-based and game-based training. They compared that to a group that, in addition, also received classroom teaching. No specific intervention was found more effective than the other, and the additional classroom training did not increase the effectiveness. Burns et al. (2019) [7] held a spear-phishing campaign within an organisation. In several phishing rounds, they measured the effect of various framing messages. The messages warned about the consequences of being phished for the individual and the organisation. They found that, although not significantly, framing messages focussing on individual loss are most effective. Rastenis et al. (2019) [19] held security training at a technical university and measured the effectiveness by sending generic phishing emails to employees. Trained subjects scored up to ten times better. Ultimately, Weaver et al. (2021) [25] trained people within a university. They used an online Jigsaw phishing quiz to train the subjects. The subjects receiving the training did a better job in recognising phishing emails than the control group.

All of the studies mentioned above share the same characteristics.

1. The context of the attack; what social engineering attack was used, and whether subjects were pre-victimised before receiving an intervention.
2. The characteristics of the intervention; modality, priming, warning, focus, format, tips and intensity.
3. The characteristics of the evaluation study; the environment of the experiment, the time between intervention and experiment, awareness of the subjects and randomisation of subjects.

4. METHODOLOGY

The methodology consists of four parts. First, relevant literature will be collected and classified. We will do this in the same approach as Bullée and Junger (2020) [5]. Then, a meta-analysis will be performed. The meta-analysis should find how effective the current interventions are (SQ1). Afterwards, a subgroup analysis will be done to determine what aspects of interventions are effective (SQ2). Finally, a regression analysis will be done on the effectiveness of interventions and the time between intervention and experiment. This will show the impact of time on the efficacy of interventions (SQ3).

4.1 Study selection

4.1.1 Scopus query

To retrieve our initial set of articles, we performed a query on the Scopus database. The query that was used is: '*(KEY (("social engineering") OR (phishing) OR ((disclosure) AND ((online) OR (cybercrime) OR (internet))) AND ((experiment*) OR (training) OR (survey) OR (warning) OR (intervention)))) AND (EXCLUDE (SUBJAREA, "MEDI"))*'

To include as many studies as possible within the available time for this research, we included all studies from 2015 till 2021 that were not excluded by the query, in total, that were 439 studies.

4.1.2 Exclusion criteria

For this meta-analysis, we looked for studies that test the effectiveness of a social engineering intervention and use an experimental design with human subjects. To determine whether to exclude a study from our meta-analysis, we use the same criteria as Bullée and Junger (2020) [5]. A summary of the criteria can be found below.

1. The study is a scientific paper or PhD thesis.
2. The language must be English or Dutch.
3. The study involves human subjects.
4. An experimental design must be used.
5. The experiment should aim to reduce victimisation.
6. There must be a comparison of at least two groups.
7. No technical solutions are used.
8. There should be at least 20 observations per group.

4.1.3 Abstract screening

All 439 studies went through the abstract screening. In this phase, titles and abstracts of the articles were read to exclude the studies that clearly met one or more of the exclusion criteria. In total, 392 studies were excluded in this phase. The most common reasons for exclusion were that studies were not related to social engineering or did not aim to reduce victimisation. Other common reasons for exclusion were lacking an experimental design and proposing a technical solution.

4.1.4 Full article screening

In this phase, from the 47 studies, the entire article was obtained, and their methodology section was read to determine if the study passed all eight criteria. 31 studies were excluded, and 15 studies were included in the meta-analysis. Common reasons for exclusion were not aiming to reduce victimisation, not comparing two groups of subjects, and offering technical solutions.

4.1.5 Independent variables

Studies use different methodologies to measure the effectiveness of various social engineering interventions. The categorisation scheme of Bullée and Junger (2020)[5] was used to compare the efficacy of multiple aspects of interventions. The categorisation consists of the three broad categories mentioned in section 3: the context of the social engineering attack, the characteristics of interventions and the characteristics of the evaluation study.

For the context of the attack, the following two coding categories were used:

- Type of social engineering: Measurement of the device which was targeted:
  1) Email (e.g., phishing email)
  2) Face-to-face (e.g., asking for credentials in person)
  3) Phone (e.g., asking for authorizations codes through during a phone call)
  4) Website (e.g., a login screen on a fraudulent website)
- Pre-victimisation: Whether the target previously fell victim to a social engineering attack in the study.
  1) No, every subject received the intervention.
  2) Yes, only the victimised subjects receive the intervention.

For the characteristic of the intervention, the following seven coding categories were used:

- Modality: Measures what device was used to give the intervention.
  1) Dynamic content (e.g., games and quizzes)
  2) Orally (e.g., Lectures or videos)
  3) Static content (e.g., Documents)
- Priming: Mentioning certain words (e.g., phishing, cybercrime) can activate knowledge.
  1) No
  2) Yes
- Warning: Whether the subjects were warned about social engineering.
  1) No, there was no warning given
  2) Yes, a warning was given
  3) Yes, a warning was given in combination with training materials
- Focus: Measures what the intervention was focussing on.
  1) Cybercrime, cybercrime in general
  2) Email, how to recognise fraudulent emails
  3) URL, the structure of URLs and how to recognise a fraudulent URL
  4) URL and Email, both recognising fraudulent emails as fraudulent URLs
  5) Social Engineering, social engineering in general
  6) Other, Another focus or a combination of the above
- Format: Measures the form of the intervention
  1) Text, the content of the intervention was presented in a textual way only.
  2) Text and graphics, the content of the intervention was presented with text and graphics.
  3) Comic, a comic (in combination with text) was used to present the intervention.
  4) Game, the intervention consisted of a game or quiz
  5) Other, another format was used, or different combinations not mentioned above were used.
- Tips: The subjects were given tips (e.g., how to recognise a fraudulent URL).
  1) No, tips were not given
  2) Yes, tips were given
  3) Tips and training, tips along with training material or during training were given.
- Intensity:
  1) Low (e.g., pamphlets)
  2) Medium (e.g., long reading materials or short videos)
  3) High (e.g., games and lectures)

For the characteristic of the evaluation study, the following four coding categories were used:

- Delay: The delay in hours between the intervention and the test.
- Environment: Whether the environment was in a laboratory where subjects were observed or in a more natural 'real' environment.
  1) Lab, the subjects were tested in a laboratory setting
  2) Real, the subjects were tested in everyday life conditions
- Aware: Whether subjects were aware of their participation in an experiment.
  1) No, subjects were not aware of their participation.
  2) Yes, subjects were aware of their involvement and the purpose of the study
  3) Semi, subjects were aware of their participation but unaware of the intention of the study
- Randomisation:
  1) Not applicable, there was no randomisation or randomisation was not mentioned
  2) Quasi, a quasi-experiment was used (No randomisation)
  3) Yes, subjects were randomised

4.1.6 Dependent variable

For the dependent variable, we measure the effect size of the intervention. We use Cohen's d or standardised mean difference (SMD) to measure the effect size. SMD is the difference between the mean of the intervention and control group, divided by the pooled standard deviation of both groups [9]. An SMD of 0.2 is considered a low effect size. An SMD of 0.5 is regarded as a medium effect size, and an SMD of 0.8 is regarded as a high effect size [9].

4.2 Meta-analysis

After the classification and calculation of the effect sizes are finished, a meta-analysis will be performed. For this study, we used the software of Comprehensive Meta-Analysis (CMA) [10]. This study uses a random-effects model to estimate the mean effect size. Attack vectors, interventions and methodological approaches vary among studies, so we cannot assume a fixed effect for all interventions.

4.3 Subgroup analysis

Then, a subgroup analysis was done. Interventions were grouped by all characteristics using CMA [10]. CMA allows to group the results of the meta-analysis by categorical variables. This was done for all categories except delay. The subgroup analysis shows if specific characteristics of interventions are significantly more effective than others.

4.4 Regression analysis

Ultimately, after conducting the meta-analysis, we perform a regression analysis. As an independent variable, we take the time between intervention and measurement. We again use the effect size (SMD) as a dependent variable.

5. RESULTS

The results of the meta-analysis will be presented below. In total, 15 different studies were included in the meta-analysis. These studies consisted of a total of 27 measurements and 16.670 participants. First, the overall results of the meta-analysis will be discussed. Then we will look into the results of the subgroup analysis and see what aspects of the interventions are significantly more effective than others. Lastly, we will look at the effect of time on the effectiveness of interventions.

5.1 Overall results

In Figure 1, the effectiveness of the intervention(s) in the selected studies can be viewed as well as their 95% confidence interval. An SMD larger than zero indicates a positive effect, whereas an SMD smaller than zero indicates a negative effect [7]. The pooled effect size of all 27 measurements in the meta-analysis suggest a medium effectiveness for social engineering interventions (SMD = 0,503, CI = [0,375; 0,720]).

5.2 Subgroup analysis

For the independent variables, a subgroup analysis was made. The results can be viewed in Table 1.

5.2.1 Type of social engineering

Interventions tested by phone (SMD = 1.370) or website (SMD = 1.399) were effective in reducing victimisation. Interventions tested by email (SMD = 0.321) had a low effect. Face to face testing had a negative effect (SMD = -0.239). However, interventions tested by phone and face to face had only a few measurements (n = 1 and 2 respectively). The type of social engineering was statistically significant (*p* = .004)

5.2.2 Pre-victimised

Interventions that pre-victimised the subjects had a low effect on reducing victimisation (SMD = 0.232). Interventions, where no pre-victimisation was used, had a medium effect size (SMD = 0.659). Pre-victimisation was statistically significant (*p*=.000).

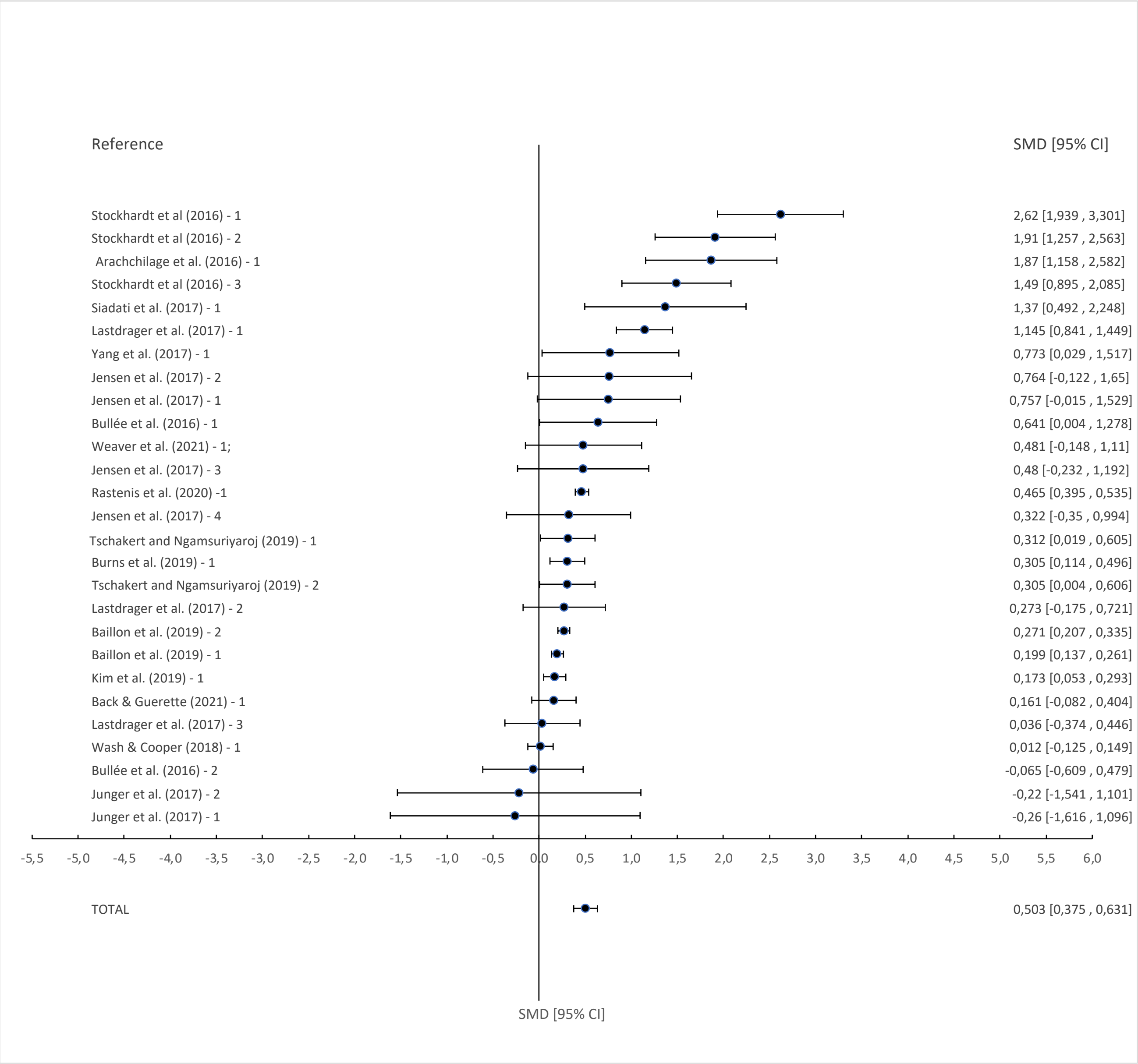

Figure 1. The effect sizes (SMD) and 95% CI of the individual measurements and the pooled measurement

5.2.3 *Modality*

Interventions delivered through static content had a low effect size (SMD = 0.285). Orally delivering the intervention had a medium effect (SMD = 0.621) and Interventions delivered through dynamic content had a high effect size (SMD = 1.112). The modality was statistically relevant (*p* = .023).

5.2.4 *Priming*

Interventions that used priming had a low effect size (SMD = 0.430), whereas interventions that did not use priming had a medium effect size (SMD = 0.619). The use of priming was statistically not significant (*p* = .326).

Table 1. Average effect sizes and 95% CI of the subcategories.

Characteristic	**Type**	**SMD**	**95% CI**	**n**	**I**	**p**
All		0,503	[0,375 ; 0,631]	27	87,15	-
Context						
Type of social engineering						.004
	Face to Face	-0,239	[-1,186 ; 0,708]	2	0,00	
	Email	0,321	[0,053 ; 0,589]	18	81,05	
	Telephone	1,370	[0,492 ; 2,248]	1	0,00	
	Website	1,399	[0,591 ; 2,207]	6	89,78	
Pre-victimised						.000
	No	0,659	[0,476 ; 0,842]	21	88,96	
	Yes	0,232	[0,107 ; 0,357]	6	60,93	
Characteristics of the intervention						
Modality						.023
	Orally	0,621	[0,337 ; 0,905]	8	91,77	
	Static	0,285	[0,165 ; 0,405]	15	68,24	
	Dynamic	1,112	[0,241 ; 1,983]	4	90,25	
Priming						.326
	No	0,619	[0,267 ; 0,971]	12	75,53	
	Yes	0,430	[0,291 ; 0,569]	15	90,00	
Warning						.000
	No	0,189	[-0,352 ; 0,730]	3	37,33	
	Yes	0,223	[0,139 ; 0,307]	10	60,43	
	Warning + train	0,867	[0,561 ; 1,173]	14	88,45	
Focus						.000
	URL	1,739	[1,176 ; 2,302]	5	71,41	
	Email	0,465	[0,395 ; 0,535]	1	0,00	
	URL+Email	0,308	[0,174 ; 0,442]	11	79,78	
	Social engineering	0,272	[0,048 ; 0,496]	4	11,52	
	Cybercrime	0,193	[0,069 ; 0,317]	1	5,60	
	Other	1,370	[0,492 ; 2,248]	5	0,00	
Format						.654
	Text	0,577	[-0,014 ; 1,168]	7	82,63	
	Comic	0,283	[0,008 ; 0,558]	3	29,15	
	Game	0,482	[0,075 ; 0,889]	3	0,00	
	Text + comic	0,596	[0,332 ; 0,860]	4	93,71	
	Other	0,525	[0,294 ; 0,756]	10	89,96	
Tips						.000
	No	0,640	[-0,948 ; 2,228]	2	74,43	
	Yes	0,201	[0,120 ; 0,282]	9	55,83	
	Yes + train	0,771	[0,518 ; 1,024]	16	87,06	
Intensity						.001
	Low	0,209	[0,098 ; 0,320]	7	69,95	
	Medium	0,634	[0,274 ; 0,994]	8	55,09	
	High	0,684	[0,439 ; 0,929]	12	91,19	
Characteristics of testing methods						
Retention of knowledge		β = -0,00034				.097
Environment						.008
	Lab	1,194	[0,557 ; 1,831]	8	82,00	
	Real	0,321	[0,216 ; 0,426]	19	81,32	
Aware						.015
	No	0,260	[0,142 ; 0,378]	9	86,65	
	Half	1,108	[0,539 ; 1,677]	10	88,92	
	Yes	0,372	[0,044 ; 0,7]	8	75,36	
Randomisation						.289
	No	0,449	[0,383 ; 0,515]	3	0,00	
	Quasi	-0,239	[-1,186 ; 0,708]	2	0,00	
	Yes	0,546	[0,391 ; 0,701]	20	87,62	
	N/A	0,981	[-0,692 ; 2,654]	2	94,96	

5.2.5 Warning

Using and not using warnings had a low effect size (SMD = 0.223 and 0.189 respectively). However, using warnings combined with training materials did have a high effect size (SMD = 0.867). Using warnings with training materials was not statistically significant (*p* = .0004).

5.2.6 Focus

Interventions focussing on recognising fake URLs had a very high effect size (SMD = 1.739). Interventions focussing on email, URL and email, social engineering or cybercrime had low effect sizes (SMD = 0.465, 0.308, 0.272 and 0.193 respectively). Interventions that could not be classified in any of these groups, had a very high effect size (SMD = 1.370). Interventions that could not be classified did not present a new focus group but used a combination of cybercrime, social engineering, email and URL in their interventions. The focus of the intervention was statistically significantly (*p* = .0004).

5.2.7 Format

Interventions with a textual format had a medium effect size (SMD = 0.577). Interventions that used a comic had a low effect size (SMD = 0.283). A combination of text and comic had a medium effect size (SMD = 0.596). The usage of a game had a medium/low effect size (SMD = 0.482). Interventions that used a different format (e.g., presentation) or a combination of formats had a medium effect size (SMD = 0.525). The format of the intervention was not statistically significant (*p* = .564).

5.2.8 Tips

Not using tips had a medium effect size (SMD = 0.640) whereas only using tips had a low effect size (SMD = 0.201). Using tips in combination with training materials had a medium/high effect rate (SMD = 0.771). Using tips in combination with training material was statistically significant (*p* = .0001).

5.2.9 Intensity

Interventions with a low intensity had a low effect size (SMD = 0.209). Interventions with a medium or high intensity had a medium effect size (SMD = 0.634 and 0.684 respectively). The intensity of the intervention was statistically relevant (*p* = .0005)

5.2.10 Environment

Interventions tested in a laboratory setting had a high effect size (SMD = 1.108), whereas interventions tested in a natural environment had a low effect size (SMD = 0.321). The environment where the intervention was tested was statistically significant (*p* = .008).

5.2.11 Awareness

Interventions, where subjects were unaware that they participated in an experiment had a low effect size (SMD = 0.260). Interventions, where subjects knew about their participation but didn't know about the purpose, had a high effect size (SMD = 1.108). Interventions where subjects were both aware of the purpose, and th eir participation had a low effect size (SMD = 0.372). The awareness of subjects was statistically significant (*p* = .015).

5.2.12 Randomisation

Studies that used randomisation of subjects across groups did have a medium effect size (SMD = 0.546). Studies that did not use randomisation had a lower effect size (SMD = 0.449). Studies that used a Quasi-experimental design had a negative effect size (SMD = -0.239). Randomisation of subjects was not statistically significant (p = .289).

5.3 Regression analysis

The regression analysis showed a small decrease in effect size of 0.00034 per hour (*p* = .097). Figure 2 shows a plot of the regression analysis.

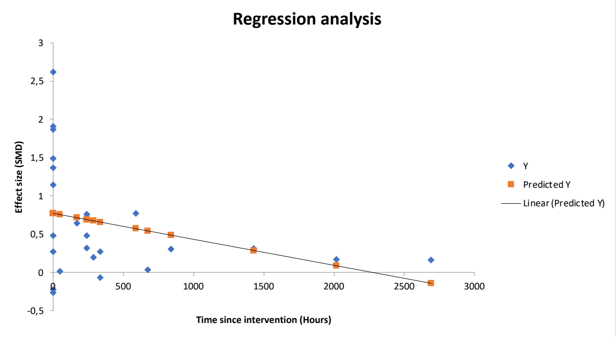

Figure 2: The effect of time on the effect size of interventions

6. DISCUSSION

In this section, we will discuss the results of section 5. Overall, we see that interventions do have a medium effect size. Some aspects of interventions were more effective than others. Finally, we observed that time has no significant impact on the effect size.

6.1 Meta-analysis

The results of the meta-analysis show that social engineering interventions do have a medium effect size (SMD = 0.503) [9]. This indicates that interventions can reduce the effectiveness of social engineering attacks. The result is in line with the findings of Bullée and Junger (2020) [5]. This finding answers SQ1.

6.2 Subgroup analysis

The results of the subgroup analysis showed that interventions were not effective against all types of social engineering attacks. Face to face attacks had a small negative effect size, meaning that the interventions do more harm than good against these types of attacks. However, this could be explained by the limited number of face to face attacks in this study. Interventions were also less effective against attacks that use emails to victimise subjects. This can be because most experiments using email as the attack vector were using a real environment and had a delay between the intervention and the attack. The interventions were effective against website-based and phone-based attacks. Victimising subjects before the intervention did not show a statistically relevant difference.

Some aspects of interventions are more effective than others. However, for two aspects, no difference was found. According to the results, it does not matter what format is used. This is in line with the results of Bullée and Junger (2020) [5]. In addition, the usage of priming had no significant effect. This result is in line with [13] but in contrast with [5].

The other five characteristics of interventions did show a significant difference in effect size. For the modality of interventions, dynamic content such as videos, games and quizzes were most effective followed by oral content such as presentations. Textual content had a low effect size. These results are similar to the findings of [5], with the difference that dynamic content shows a slightly higher effect size than [5]. The use of only warnings had a low effect size. Not using warnings had a slightly lower effect. However, the usage of warnings in combination with training materials had a high effect size. This can be explained by the usage of extra training materials. Giving subjects additional training materials shows a better performance [17]. The finding that warnings alone are not effective is in line with [13]. For the focus of the intervention, a focus on URLs appears to be very effective. Focussing on emails or emails in combination with URLs were somewhat effective but not close to studies that only focussed on URLs. Focussing on only social engineering or cybercrime did not seem to be effective. Interventions that had another focus were also very effective. Interventions with an ‘other’ classification were mainly a combination of URL and email focus in combination with cybercrime or social engineering. Their effectiveness could be explained by the intensity of the intervention. The finding that URL focussed interventions are most efficient is in line with [5]. The use of tips in interventions had a low effect size but using tips in combination with additional training materials showed a high effect size. This could once again be explained by using additional training materials like we have seen for the use of warnings. Interventions that did not use warnings had a medium effect size, but that might be because our study only contained two measurements without the use of tips. Interventions with a low intensity did have a small effect size. Studies with a medium or high effect size were equally effective.

For the characteristics of testing methods, environment and awareness were statistically significant. Subjects tested in a lab environment performed better than subjects in a natural setting. This can be explained by findings that people observed in a laboratory setting can produce higher effects sizes [8]. Subjects aware of their participation in a study showed a higher effect size than subjects who did not know about participation at all. Randomising subjects across groups was not statistically significant.

With the findings above, we can combine the aspects of interventions that appear to be more effective than others. By doing so, we can answer SQ2 and potentially design more efficient interventions. The modality of interventions should be either dynamic or orally. Dynamic modality based on user input (e.g., games and quizzes) show the highest effectiveness. The interventions should include warnings and tips in combination with training materials. Without the use of training materials, warnings and tips are not efficient. The intervention should focus on recognizing fraudulent URL’s and should be of medium or high intensity. The format and use of priming are not relevant for the effect size.

6.3 Regression analysis

The results of the regression analysis show that time since the intervention is not statistically significant (p = .097), so we cannot say with 95% confidence that time influences the effect size. The decay found was 0.00034. This is lower than the findings of [5]. If statistically significant, the results would indicate that on average, interventions lose their effectiveness in approximately two months. The low significance level is unexpected. We expected to find a clear relationship between time and effectiveness. This can be explained by publication bias, where studies that find a positive result are more likely to publish than studies that find low or negative results. Linear regression assumes that extreme values are distributed evenly. This was not the case for our dataset. Unfortunately, this study was not able to investigate the decay of grouped characteristics because of the limited number of measurements in the study. With the findings in this section, we have an answer to SQ3.

7. CONCLUSION

The current social engineering interventions do have a medium effect in preventing social engineering attacks to succeed. However, some interventions are very effective, while others have no effect or a negative effect. Based on this research, we can combine the characteristics that were effective and build a blueprint for future interventions. Interventions should contain the following elements: The modality should be dynamic. Tips and warnings should be used in combination with training materials. The focus of the intervention should be to recognise fraudulent URL’s and should be of medium or high intensity.

We did observe a decrease in effectiveness when time goes by, however, the significance was not high enough to conclude with 95% confidence that time has an influence on the effect size.

8. LIMITATIONS

This study contains two limitations.

First, the study selection and categorisation were performed by only one researcher. The outcome of the classification is therefore not as scientifically sound as classifications coded by multiple researchers.

Second, the number of included studies is low. Some categories (Type: phone; Focus: email, cybercrime; Tips: no) only had one or two measurements. No real conclusions can be drawn from that.

9. FUTURE WORK

For future work, we have several suggestions:

Other suggested work is to measure the time effect of interventions by setting up a big field experiment. This would allow researchers to keep all variables equal.

Another future research could be to investigate the effect of re-training. We have seen that the effectiveness of interventions decreases over time, but we don’t know yet what the impact of re-training is.

10. PRACTICAL IMPLICATIONS

The findings of this study can be used by practitioners to make more effective interventions. Especially learning humans on how to recognise fraudulent URLs seems to be effective. Using tips and warnings in combination with training materials was also found effective. These aspects could be combined in an intervention that teaches how to recognise fraudulent URL’s with tips and warnings, and then gives the participants training materials to practice on. Training materials could be a test to identify fraudulent URL’s.

11. REFERENCES

[1] Anti-Phishing Working Group. Phishing activity trends report. In APWG (ed.), Anti Phishing Working Group, 2021. (Accessed on January 27, 2022) Available at: https://docs.apwg.org/reports/apwg_trends_report_q3_2021.pdf

[2] Arachchilage, N.A.G., Love, S. and Beznosov, K. (2016), “Phishing threat avoidance behaviour: an empirical investigation”, Computers in Human Behavior, Vol. 60, pp. 185-197. https://doi.org/10.1016/j.chb.2016.02.065

[3] Baillon, A., de Bruin, J., Emirmahmutoglu, A., van de Veer, E., & van Dijk, B. (2019). Informing, simulating experience, or both: A field experiment on phishing risks. *PLOS ONE*, *14*(12), e0224216. https://doi.org/10.1371/journal.pone.0224216

[4] Bakhshi, T. (2017). Social engineering: Revisiting end-user awareness and susceptibility to classic attack vectors. *2017 13th International Conference on Emerging Technologies (ICET)*. Published. https://doi.org/10.1109/icet.2017.8281653

[5] Bullee, J. W., & Junger, M. (2020). How effective are social engineering interventions? A meta-analysis.

Information & Computer Security, *28*(5), 801–830. https://doi.org/10.1108/ics-07-2019-0078

[6] Bullée, J.H., Montoya, L., Junger, M. and Hartel, P.H. (2016), "Telephone-based social engineering attacks: an experiment testing the success and time decay of an intervention", Cryptology and Information Security Series, Vol. 14, pp. 107-114. https://doi.org/10.3233/978-1-61499-617-0-107

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

[8] Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2016), "Evaluating replicability of laboratory experiments in economics", Science, Vol. 351 No. 6280, pp. 1433-1436, available at: https://doi.org/10.1126/science.aaf0918

[9] Cohen, J. (2013), "Statistical power analysis for the behavioral sciences", Retrieved from https://books.google.nl/books?id=cIJH0lR33bgC

[10] Comprehensive Meta-Analysis. (2022, January 1). *Take a Tour of Comprehensive Meta-Analysis*. Biostat. Retrieved January 21, 2022, from https://www.meta-analysis.com/pages/comprehensive_meta-analysis_tour.php?cart=BHMA6302295

[11] IBM Security. (2021). *Cost of a Data Breach Report 2021*. https://www.dataendure.com/wp-content/uploads/2021_Cost_of_a_Data_Breach_-2.pdf

[12] Jensen, M.L., Dinger, M., Wright, R.T. and Thatcher, J.B. (2017), "Training to mitigate phishing attacks using mindfulness techniques", Journal of Management Information Systems, Vol. 34 No. 2, pp. 597-626. https://doi.org/10.1080/07421222.2017.1334499

[13] Junger, M., Montoya, L. and Overink, F.-J. (2017), "Priming and warnings are not effective to prevent social engineering attacks", Computers in Human Behavior, Vol. 66, pp. 75-87, available at: https://doi.org/10.1016/j.chb.2016.09.012

[14] Kim, B., Lee, D. Y., & Kim, B. (2019). Deterrent effects of punishment and training on insider security threats: a field experiment on phishing attacks. *Behaviour & Information Technology*, *39*(11), 1156–1175. https://doi.org/10.1080/0144929x.2019.1653992

[15] Lastdrager, E.E., Carvajal Gallardo, I., Hartel, P.H. and Junger, M. (2017), "How effective is antiphishing training for children?", Thirteenth Symposium on Usable Privacy and Security (Soups 2017). Santa Clara, CA: USENIX Association.

[16] Mouton, F., Leenen, L., & Venter, H. (2016). Social engineering attack examples, templates and scenarios. *Computers & Security*, *59*, 186–209. https://doi.org/10.1016/j.cose.2016.03.004

[17] Nguyen, C., Jensen, M., & Day, E. (2021). Learning not to take the bait: a longitudinal examination of digital training methods and overlearning on phishing susceptibility. *European Journal of Information Systems*, 1–25. https://doi.org/10.1080/0960085x.2021.1931494

[18] Parthy, P. P., & Rajendran, G. (2019). Identification and prevention of social engineering attacks on an enterprise. *2019 International Carnahan Conference on Security Technology (ICCST)*. Published. https://doi.org/10.1109/ccst.2019.8888441

[19] Rastenis, J., Ramanauskaitė, S., Janulevičius, J., & ČEnys, A. (2020). Impact of Information Security Training on Recognition of Phishing Attacks: A Case Study of Vilnius Gediminas Technical University. *Communications in Computer and Information Science*, 311–324. https://doi.org/10.1007/978-3-030-57672-1_23

[20] Siadati, H., Nguyen, T., Gupta, P., Jakobsson, M. and Memon, N. (2017), "Mind your smses: mitigating social engineering in second factor authentication", Computers and Security, Vol. 65, pp. 14-28. https://doi.org/10.1016/j.cose.2016.09.009

[21] Stockhardt, S., Reinheimer, B., Volkamer, M., Mayer, P., Kunz, A., Rack, P. and Lehmann, D. (2016), "Teaching phishing-security: which way is best?", IFIP Advances in Information and Communication Technology, Vol. 471, pp. 135-149, available at: https://doi.org/10.1007/978-3- 319-33630-5_10

[22] Tschakert, K. F., & Ngamsuriyaroj, S. (2019). Effectiveness of and user preferences for security awareness training methodologies. *Heliyon*, *5*(6), e02010. https://doi.org/10.1016/j.heliyon.2019.e02010

[23] Wang, Z., Sun, L., & Zhu, H. (2020). Defining Social Engineering in Cybersecurity. *IEEE Access*, *8*, 85094–85115. https://doi.org/10.1109/access.2020.2992807

[24] Wash, R., & Cooper, M. M. (2018). Who Provides Phishing Training? *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3173574.3174066

[25] Weaver, B. W., Braly, A. M., & Lane, D. M. (2021). Training Users to Identify Phishing Emails. *Journal of Educational Computing Research*, *59*(6), 1169–1183. https://doi.org/10.1177/0735633121992516

[26] Yang, W., Xiong, A., Chen, J., Proctor, R. W., & Li, N. (2017). Use of Phishing Training to Improve Security Warning Compliance. *Proceedings of the Hot Topics in Science of Security: Symposium and Bootcamp on - HoTSoS*. https://doi.org/10.1145/3055305.3055310

APPENDIX A. CLASSIFICATION

reference - measurement	N	SMD	Delay	Type	PreVic	Modality	Priming	Warning	Focus	Format	Tips	Intensity	Environ	Aware	Random
Arachchilage et al. (2016) - 1	20	1,87	0	Web	No	Dynamic	No	Yes + t	URL	Text + Comic	Yes + training	High	Lab	Half	N/A
Back & Guerette (2021) - 1	384	0,16	2688	Mail	No	Static	Yes	Yes + t	Cybercrime	Other	Yes + training	High	Real	Half	N/A
Baillon et al. (2019) - 1	4006	0,2	288	Mail	No	Static	Yes	Yes	URL + Email	Text + Comic	Yes	Low	Real	No	Yes
Baillon et al. (2019) - 2	3805	0,27	0	Mail	Yes	Static	Yes	Yes	URL + Email	Text + Comic	Yes	Low	Real	No	Yes
Bullée et al. (2016) - 1	64	0,64	168	Web	No	Static	No	No	Social Engineering	Comic	Yes	Medium	Real	No	Yes
Bullée et al. (2016) - 2	63	-0,07	336	Web	No	Static	No	No	Social Engineering	Comic	Yes	Medium	Real	No	Yes
Burns et al. (2019) - 1	442	0,31	840	Mail	Yes	Static	Yes	Yes	Social Engineering	Comic	Yes	Low	Real	No	Yes
Jensen et al. (2017) - 1	73	0,76	240	Mail	No	Static	No	Yes + t	Cybercrime	Text	Yes + training	Medium	Real	Half	Yes
Jensen et al. (2017) - 2	60	0,76	240	Mail	No	Static	No	Yes + t	Cybercrime	Game	Yes + training	Medium	Real	Half	Yes
Jensen et al. (2017) - 3	67	0,48	240	Mail	No	Static	No	Yes + t	URL + Email	Text	Yes + training	Medium	Real	Half	Yes
Jensen et al. (2017) - 4	66	0,32	240	Mail	No	Static	No	Yes + t	URL + Email	Game	Yes + training	Medium	Real	Half	Yes
Junger et al. (2017) - 1	184	-0,26	0	F2F	No	Static	Yes	No	Cybercrime	Text	No	Low	Lab	Yes	Quasi
Junger et al. (2017) - 2	190	-0,22	0	F2F	No	Static	No	Yes	Social Engineering	Text	Yes	Low	Lab	Yes	Quasi
Kim et al. (2019) - 1	1147	0,17	2016	Mail	No	Orally	Yes	Yes	Cybercrime	Other	Yes	High	Real	No	Yes
Lastdrager et al. (2017) - 1	189	1,15	0	Mail	No	Orally	No	Yes + t	URL + Email	Other	Yes + training	High	Real	Yes	Yes
Lastdrager et al. (2017) - 2	81	0,27	336	Mail	No	Orally	No	Yes + t	URL + Email	Other	Yes + training	High	Real	Yes	Yes
Lastdrager et al. (2017) - 3	96	0,04	672	Mail	No	Orally	No	Yes + t	URL + Email	Other	Yes + training	High	Real	Yes	Yes
Rastenis et al. (2020) -1	3266	0,47	8760	Mail	No	Orally	Yes	Yes + t	Email	Other	Yes + training	High	Real	No	No
Tschakert and Ngamsuriyaroj (2019) - 1	50	0,31	1428	Mail	Yes	Dynamic	Yes	Yes	URL + Email	Other	Yes + training	High	Real	Yes	No
Tschakert and Ngamsuriyaroj (2019) - 2	49	0,31	1428	Mail	Yes	Orally	Yes	Yes	URL + Email	Other	Yes + training	High	Real	Yes	No
Wash & Cooper (2018) - 1	2165	0,01	48	Mail	Yes	Static	Yes	Yes	URL + Email	Text	Yes	Low	Real	No	Yes
Weaver et al. (2021) - 1	40	0,48	0	Mail	Yes	Dynamic	Yes	Yes + t	URL + Email	Game	Yes + training	High	Lab	Yes	Yes
Yang et al. (2017) - 1	30	0,77	588	Mail	No	Orally	Yes	Yes	URL	Other	Yes	Medium	Lab	Half	Yes
Siadati et al. (2017) - 1	52	1,37	0	Phone	No	Static	No	Yes	Other	Text	No	Low	Real	No	Yes
Stockhardt et al (2016) - 1	30	2,62	0	Web	No	Orally	Yes	Yes + t	URL	Other	Yes + training	High	Lab	Half	Yes
Stockhardt et al (2016) - 2	25	1,91	0	Web	No	Dynamic	Yes	Yes + t	URL	Text + Comic	Yes + training	High	Lab	Half	Yes
Stockhardt et al (2016) - 3	26	1,49	0	Web	No	Static	Yes	Yes + t	URL	Text	Yes + training	Medium	Lab	Half	Yes