How effective are social engineering interventions? A meta-analysis

Article *in* Information and Computer Security · August 2020

DOI: 10.1108/ICS-07-2019-0078

CITATIONS
0

READS
127

2 authors:

Jan-Willem H. Bullée

17 PUBLICATIONS **103** CITATIONS

SEE PROFILE

Marianne Junger

University of Twente

223 PUBLICATIONS **2,841** CITATIONS

SEE PROFILE

How effective are social engineering interventions? A meta-analysis

Jan-Willem Bullee and Marianne Junger
Department of Industrial Engineering and Business Information Systems (IEBIS), Faculty of Behavioural, Management and Social Sciences (BMS), University of Twente, Enschede, The Netherlands

Received 2 July 2019
Revised 26 February 2020
27 May 2020
Accepted 27 May 2020

Abstract

Purpose – Social engineering is a prominent aspect of online crime. Various interventions have been developed to reduce the success of this type of attacks. This paper aims to investigate if interventions can help to decrease the vulnerability to social engineering attacks. If they help, the authors investigate which forms of interventions and specific elements constitute success.

Design/methodology/approach – The authors selected studies which had an experimental design and rigorously tested at least one intervention that aimed to reduce the vulnerability to social engineering. The studies were primarily identified from querying the Scopus database. The authors identified 19 studies which lead to the identification of 37 effect sizes, based on a total sample of N = 23,146 subjects. The available training, intervention materials and effect sizes were analysed. The authors collected information on the context of the intervention, the characteristics of the intervention and the characteristics of the research methodology. All analyses were performed using random-effects models, and heterogeneity was quantified.

Findings – The authors find substantial differences in effect size for the different interventions. Some interventions are highly effective; others have no effect at all. Highly intensive interventions are more effective than those that are low on intensity. Furthermore, interventions with a narrow focus are more effective than those with a broad focus.

Practical implications – The results of this study show differences in effect for different elements of interventions. This allows practitioners to review their awareness campaigns and tailor them to increase their success.

Originality/value – The authors believe that this is the first study that compares the impact of social engineering interventions systematically.

Keywords Awareness, Cybercrime, Intervention, Meta-analysis, Online, Phishing, Social engineering, Systematic review

Paper type Literature review

Introduction

Many statistics show that online crime has become one of the major threats for both individuals (Henson *et al.*, 2016; Internet Crime Complaint Center, 2018; Marinos and Sfakianakis, 2012; Reep-van den Bergh and Junger, 2018) and organisations (Klahr *et al.*, 2017). Many of these online crimes contain an element of fraud and deception, or "social engineering" (Blakeborough and Correia, 2017; Verizon Risk Team, 2018) to achieve their goal.

Social engineering is a non-technical type of attack based on human interaction and complements technical attacks. Abraham and Chengalur-Smith (2010) proposed the following definition:

Information & Computer Security

2056-4961
DOI 10.1108/ICS-07-2019-0078

> The use of social disguises, cultural ploys, and psychological tricks to get computer users (i.e. targets) to assist hackers (i.e. offenders) in their illegal intrusion or use of computer systems and networks.

The social engineering attack vector is considered one of the biggest threat to information systems (Rouse, 2006). One of the dangers of social engineering attacks is their harmless and legitimate appearance so that targets are unaware of being victimised (Hadnagy and Wilson, 2010; The Federal Bureau of Investigation, 2015). Deception and manipulation are used by offenders to make targets assist in their victimisation (Bosworth *et al.*, 2014).

Cybersecurity incidents are more often caused by human failure (Chan *et al.*, 2005) than by failing technology (Schneier, 2000b). Humans are still considered as "the weakest link", in security (Camp *et al.*, 2019; Glaspie and Karwowski, 2018; Happ *et al.*, 2016; Heartfield and Loukas, 2018; Parsons *et al.*, 2017; Rouse, 2006; Schneier, 2000b). To put it bluntly: "Only amateurs attack machines, professionals attack humans" (Schneier, 2000a).

Researchers have developed interventions to prevent social engineering and many reported positive results. For instance, several anti-phishing training programs were found to be effective (Dodge *et al.*, 2007; Kumaraguru *et al.*, 2009; Kumaraguru *et al.*, 2010; Sheng *et al.*, 2007). However, some interventions found no effect of training (Davinson and Sillence, 2010) and some field replications of a previously effective intervention did not find a positive impact in contrast to the original (Caputo *et al.*, 2014). Finally, some studies reported adverse effects of interventions (Junger *et al.*, 2017; Kearney and Kruger, 2014; Wolff, 2016; Zhang *et al.*, 2014). Some reviews were negative about the effectiveness of interventions and argued that they had hardly any impact (Bada *et al.*, 2015; Ceesay *et al.*, 2018). For instance, Bada *et al.* (2015) stated about cybersecurity awareness campaigns: "Why do they fail to change behaviour?" Bada *et al.* (2015) and Ceesay *et al.* (2018) reported that "enterprises that do not have awareness training are doing 12% better than those that have training programs" (Ceesay *et al.*, 2018, p. 3).

The latter reviews are not encouraging about the possibility to develop interventions with a strong, effective and enduring impact on the vulnerability for social engineering. To get a better grip on the effectiveness of interventions, a comprehensive, quantitative overview is necessary. In the present study, we present a review of experimental research that evaluated interventions aiming to reduce social engineering attacks, and that allows us to compute an overall effect size as well as to study the impact of:

- the context;
- the characteristics of the intervention; and
- the characteristics of the research methodology.

Below, we present the "main characteristics of the studies and the interventions". The second section describes the methodology that was followed, the third section describes the findings, and we conclude in the fourth section with a discussion of our findings.

Main characteristics of the studies and the interventions

Two aspects of the context matter to understand the interventions:

(1) what type of social engineering is the focus of the intervention; and

(2) do the researchers "pre-victimise" their subjects, that is, do they send a social engineering attack before the intervention.

Type of social engineering

Interventions are diverse and focus on a range of social engineering attacks. Researchers usually develop mock attacks to test interventions; typically, some of their research subjects

receive an intervention and some not (the control group). Many interventions focused on phishing, and consequently, many studies used email as the modality of attack. Sometimes, face-to-face interaction is used: Junger *et al.* (2017) started a conversation at the market square to obtain bank account information. Using fake websites, mock attackers try to obtain login credentials from their targets. Stockhardt *et al.* (2016) aimed to teach people to look for signs that indicate a fake website. Social engineering is often associated with an "attacker" calling the target on the phone and ask for their password (Winkler and Dealy, 1995). Bullée *et al.* (2016) studied the success of a social engineering attack via the telephone (i.e. telephone scam). In an investigation to bypass the two-factor authentication (2FA), SMS messages were sent to persuade the receiver in forwarding their security codes (Siadati *et al.*, 2017).

Pre-victimisation

Interventions and training materials aim to increase awareness and change behaviour regarding a certain topic. However, providing intervention to someone who is already performing the desired behaviour is a waste of resources. Instead, providing the intervention only to those who need it is more efficient. Also, pre-victimisation in a mock attack serves to motivate a user: if they fell for the social engineering attack, this will motivate them to learn how to avoid this in the future. Furthermore, people show more effective learning when provided with immediate feedback after having been attacked (Schmidt and Bjork, 1992). Therefore, security researchers often use a two-stage approach:

(1) All subjects receive, e.g. a mock phishing email.

(2) Those who performed the desired behaviour (e.g. not clicking the link) are "left alone", whereas those who were victimised (e.g. clicked on the link) are redirected or invited to participate in a social engineering awareness training (Kumaraguru *et al.*, 2007).

The combination of pre-victimisation with an intervention has been called an "embedded" training or intervention. Several studies showed that this pre-victimisation was a relevant aspect of interventions in both laboratory studies (Kumaraguru *et al.*, 2009; Kumaraguru *et al.*, 2007; Kumaraguru *et al.*, 2007; Mayhorn and Nyeste, 2012; Sheng *et al.*, 2007) and real-life (Kumaraguru *et al.*, 2008).

Characteristics of the intervention

Studies that design security awareness interventions consist – among other things – of a mock attack to be able to measure the outcome: susceptibility to social engineering. The mock attack can be linked to the type of intervention. For example, warnings on a website are usually measured by observing the click-through rate of the visitors or who provides login credentials.

Modality interventions

Interventions are provided using different modalities: sometimes, a conversation was used to train the user; alternatively, a physical document was provided to transfer knowledge or online warnings to inform of potential danger. Sometimes, the interaction is dynamic, for instance, when users interact in a classroom with a trainer (Mayhorn and Nyeste, 2012). It has been suggested that using interactive anti-phishing training is a more effective way to enhance the ability of users to identify phishing URLs than the use of passive anti-phishing tutorials (Arachchilage *et al.*, 2016; Davinson and Sillence, 2010; Kumaraguru *et al.*, 2010;

Mayhorn and Nyeste, 2012; Sheng *et al.*, 2007). Accordingly, the modality of the training message or interaction is an essential aspect of interventions.

Some "training" methods consisted of sending mock phishing emails to users. However, the first human-oriented anti-phishing studies did not contain a training part (Dodge *et al.*, 2007). Instead, these studies tested the effect of a "gotcha" moment (i.e. when an employee got victimised by a mock phishing email and received a notification of being "victimised"). The idea is that employees realise how vulnerable they are and therefore act more carefully in the future. Dodge *et al.* (2007) showed that by repeatedly sending mock phishing emails, the number of victims is gradually reduced. Aburrous *et al.* (2010) reported similar results of declining phishing rates.

Priming for danger

People's behaviour can be altered when they are exposed to certain sights, words or sensations (Dolan *et al.*, 2010; Kenrick *et al.*, 2005). These stimuli prime people: they activate knowledge and makes it ready for use (Kenrick *et al.*, 2005). Priming often works outside conscious awareness (Dolan *et al.*, 2010; Kenrick *et al.*, 2005). In the physical world, a large amount of research supports the existence of priming effects (Cameron *et al.*, 2012).

Several interventions used various forms of priming. For example, Acquisti *et al.* (2012) looked at the effectiveness of raising privacy concerns on the disclosure of information online. They displayed cues to think about phishing, which consisted of some pages with pictures of phishing emails and the request to categorise them as "phishing" or "non-phishing". A decrease in disclosure after priming was reported (Acquisti *et al.*, 2012; Grazioli, 2004; Parsons *et al.*, 2015), whereas (Sundar *et al.*, 2013; Zhang *et al.*, 2014; Zhang and Xu, 2016) reported mixed findings, and no effect was found by Grazioli and Wang (2001). Overall, the results seem to be inconclusive about the impact of priming in an online context.

Warning against danger

Warnings are a more direct way to convey a message than priming. Traditional offline warnings have been successful in influencing behaviour (Argo and Main, 2004; Wogalter *et al.*, 2012). There are guidelines for adequate offline warnings, which were summarised by Wogalter *et al.* (2012). Important principles include:

- brevity, warnings should be as brief as possible; and
- design for the low-end receiver, meaning that warnings should not be directed at an "average person" but for people who for instance have lower competence, education, knowledge, the elderly or the disabled.

Warnings have been used in an online context, for instance, to warn against website (un) safety (Kirlappos and Sasse, 2012).

Warnings help users to behave more safely; however, many users do not adjust his/her behaviour when monetary rewards were at stake (Kirlappos and Sasse, 2012). A similar conclusion was drawn by (Christin *et al.*, 2011). In the context of social media disclosure, Zhang *et al.* (2014) found an increase of disclosure in the group that was presented with a warning banner. In line with this finding, Wu *et al.* (2006) found that users ignore toolbar warnings. Krol *et al.* (2012) found that 81.7% of their subjects ignored a warning when downloading a, potentially infected, PDF file. Other research also concluded that browser warnings overall did not have positive effects (Egelman *et al.*, 2008; Egelman and Schechter, 2013; Xiao and Benbasat, 2015).

Focus of the content to the intervention

The focus of interventions varies widely. Phishing emails often contain links to malicious websites. However, most users are not aware of the structure of URLs and domain names (Herzberg and Jbara, 2008). Consequently, swindlers often succeed in tricking users into clicking on these links. Accordingly, many anti-phishing games focus on recognising phishing URLs. Games as NoPhish and Anti-Phishing Phil teach users the structure of URLs and how this structure differs for legitimate URLs as compared to phishing URLs (Caputo *et al.*, 2014; Downs *et al.*, 2006; Kumaraguru *et al.*, 2009). Other anti-phishing interventions explain to users some more general characteristics of phishing emails. For instance, phishing emails are described as:

- Phishing emails often request for personal information.
- Phishing emails often contain a sense of urgency.
- Phishing emails often have a mismatch between the senders' email address in the "From" field and the company name or reply-to email mentioned in the body of the email.
- Phishing emails often contain a threat to stimulate a response.
- Phishing emails often contain misspelt words, odd spacing, or sloppy grammar.
- Phishing emails often contain links to phishing websites.
- Hovering the mouse over a link in an email will reveal the linked URL (Downs *et al.*, 2006).

A problem with applying these characteristics is that phishing emails change: they become increasingly sophisticated, and spear-phishing makes them also more challenging to recognise (Bullée *et al.*, 2017). In contrast, it is unlikely that the use of URLs will change soon; therefore, teaching about URLs will continue to be useful for a longer time.

Technical aspects of an intervention

Most interventions focused on humans as humans could disclose information or fall for an attack. However, some interventions build in technical countermeasures as an additional layer of security. Users can not circumvent these, even if they wanted to (Herzberg and Margulies, 2013).

Format to deliver the interventions

Interventions have been developed in many different formats. For instance, anti-phishing interventions were provided by giving users a text message, a comic, a combination of a comic and text, or a game. The format seems to matter. A comparison of findings showed that the comic outperformed the text and graphics intervention (Kumaraguru *et al.*, 2007).

Two large-scale real-world, real-life anti-phishing intervention studies examined the effect of embedded training: Kumaraguru *et al.* (2008) and Caputo *et al.* (2014). One study used a cartoon (Kumaraguru *et al.*, 2008), whereas the other study used text (Caputo *et al.*, 2014), the content of the message was similar. The cartoon (using few words) improved user behaviour within the company (Kumaraguru *et al.*, 2008). However, the text (using many words) did not prevent employees from being victimised by phishing (Caputo *et al.*, 2014). Games have been developed, usually as a more elaborate form of anti-phishing training. Gaming increases the motivation of users to learn (Sheng *et al.*, 2007). The positive effect of learning by gaming is confirmed in learning science (Clark and Mayer, 2016). The most tested anti-phishing game is anti-phishing phil (Arachchilage *et al.*, 2016; Davinson and

Sillence, 2010; Kumaraguru *et al.*, 2010; Mayhorn and Nyeste, 2012; Sheng *et al.*, 2007). This game teaches users to distinguish between legitimate URLs and phishing URLs. The main message of the game is to pay attention to URLs, as they are good indicators of phishing. Phil, the main character in the game, receives points when he eats legitimate worms (i.e. URLs) and points are subtracted when Phil eats bad worms. The game exists out of four rounds, and every round starts with a short tutorial providing anti-phishing advice. Additionally, the training includes examples and practice questions (Sheng *et al.*, 2007). The anti-phishing phil game was tested in several studies (Arachchilage *et al.*, 2016; Davinson and Sillence, 2010; Kumaraguru *et al.*, 2010; Mayhorn and Nyeste, 2012; Sheng *et al.*, 2007). More recently, a game was developed for smartphones (Arachchilage and Cole, 2011). Most anti-phishing experiments with games showed positive results in teaching users to identify phishing attacks. However, it is difficult to determine the exact effect of anti-phishing games in comparison to training interventions because many of the anti-phishing games were tested in small-scale pilot studies (Sheng *et al.*, 2007; Yang *et al.*, 2012).

Use of tips

The general idea of interventions is that users receive information about how to handle specific fraudulent situations. Therefore, several forms of anti-phishing interventions worked by providing tips to users or a specific recommendation. For instance, to protect against phishing users received the following tips:

- never click on links within emails;
- type in the website address into the Web browser;
- find and call the real customer service; and
- never give out personal information (Kumaraguru *et al.*, 2007, p. 5).

Intensity of the intervention

Some interventions were quite simple, and some were quite elaborate. A plausible expectation is that more intensive interventions will lead to stronger effects and have more impact on the long term. Accordingly, one might expect that more intensive interventions are better. However, these intensive interventions may be more time-consuming, more difficult to implement and more expensive. Therefore, a simple but effective intervention is overall preferable in terms of cost-effectiveness.

Characteristics of the testing method

The previous section discussed the characteristics of interventions aiming to improve general awareness. Besides intervention characteristics, the experimental methodology could have an impact on the study outcome. Below, we discuss the length of the study, whether the study was in the laboratory or the field, and whether randomisation was used.

Retention of knowledge and time delay of tests

Regardless of the design, content or quality of training, some studies showed that retaining gained knowledge is difficult for users. Studies tested retention of knowledge after 16 days (Alnajim and Munro, 2009), four weeks (Lastdrager *et al.*, 2017) or a few months (Canova *et al.*, 2015; Caputo *et al.*, 2014).

Environment: real-life or lab

By using experiments, the subjects' behaviour can be observed in a controlled environment (Siedler and Sonnenberg, 2010). Conducting experiments involving social engineering requires careful planning and consideration, e.g. how to introduce the study to subjects. People who are aware of the goal of the experiment could be biased in their behaviour. It is unlikely that they would perform the same behaviour (e.g. have similar levels of suspicion) outside of the experiment, and this could influence the ecological validity of the study (Furnell, 2007). It is therefore expected that the outcome is different for studies (in a laboratory setting, in which subjects know they are being tested) and those in the natural environment of the subject (i.e. field experiment).

Aware of being tested

Being aware of participating as a research subject relates to the observer effect. People tend to change aspects of their behaviour, when aware of being observed and could potentially influence the research outcomes (Monahan and Fisher, 2010).

Randomisation

Stronger research designs have both stronger internal and external validity (Campbell and Stanley, 1963). The use of randomised experiments is the best research design to conclude the effect of treatments (Feder *et al.*, 2000). A problem was noted by Weisburd *et al.* (2001); in a systematic review of criminal justice studies, they found that the quality of the research designs is related to the strength of the outcome of the intervention. Research designs with a weaker internal validity, are more likely to report effects in favour of the treatment group. In contrast, research designs with stronger internal validity reported weaker effects for interventions (Weisburd *et al.*, 2001). In sum, the better the design, the weaker the intervention seems to be. A recent update confirmed these findings (Welsh *et al.*, 2011).

Therefore, the present study will investigate whether the context, the characteristics of the intervention and the characteristics of the research methodology affect the outcome of anti-social engineering interventions.

Research question

The present study has two objectives: First, to investigate if interventions reduce victimisation by social engineering. Second, to examine whether the context of the experiment, intervention and study characteristics have an impact on intervention outcome.

Methods

A meta-analysis was used to answer the research questions. Below, we describe the methodology of our study.

Information sources

The Scopus database was queried to obtain studies for the analysis. The database was queried on 30 December 2017, using the following query:

> 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”))

Eligibility criteria

There were nine eligibility criteria for records to be included in the meta-analysis. Records that did not meet all eligibility criteria were excluded:

(1) To be a published scientific paper or a PhD thesis.

(2) The manuscript must be written in English or Dutch; the authors are both proficient in those two languages.

(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) There was no restriction regarding publication date.

The search query returned 348 records. Furthermore, searches were supplemented by hand searches of the reference lists of eligible studies, inquiries with colleagues and papers obtained in previous research. In this way, 70 additional references were identified, leading to a total of 418 records.

Study selection

First, the titles and abstracts of all 418 records were screened for eligibility. In this stage, 210 studies were excluded from the analysis. For the remaining 217 studies, the eligibility criteria were looked for in the full text. At this stage, 195 studies were excluded from the analysis. Reasons for exclusion were, among others, not having a control or intervention group. The remaining 22 studies were included in the qualitative synthesis. From each study, data were extracted (i.e. quantitative synthesis). For six studies, the authors were contacted since data extraction was incomplete. The published manuscript did not always present all data that was needed to perform the meta-analysis. E.g. Stockhardt *et al.* (2016) presented the results of comparing six groups in a single statistic (i.e. the outcome of an ANOVA). The authors were contacted and asked to provide the mean and standard deviation for each group. Three authors provided the required missing information, and the other three studies were excluded from the analysis. In total, 19 studies are included in the analysis. A single study can test various interventions and, accordingly, multiple effect sizes can be computed. For an overview of steps and the related number of studies, refer to Figure 1.

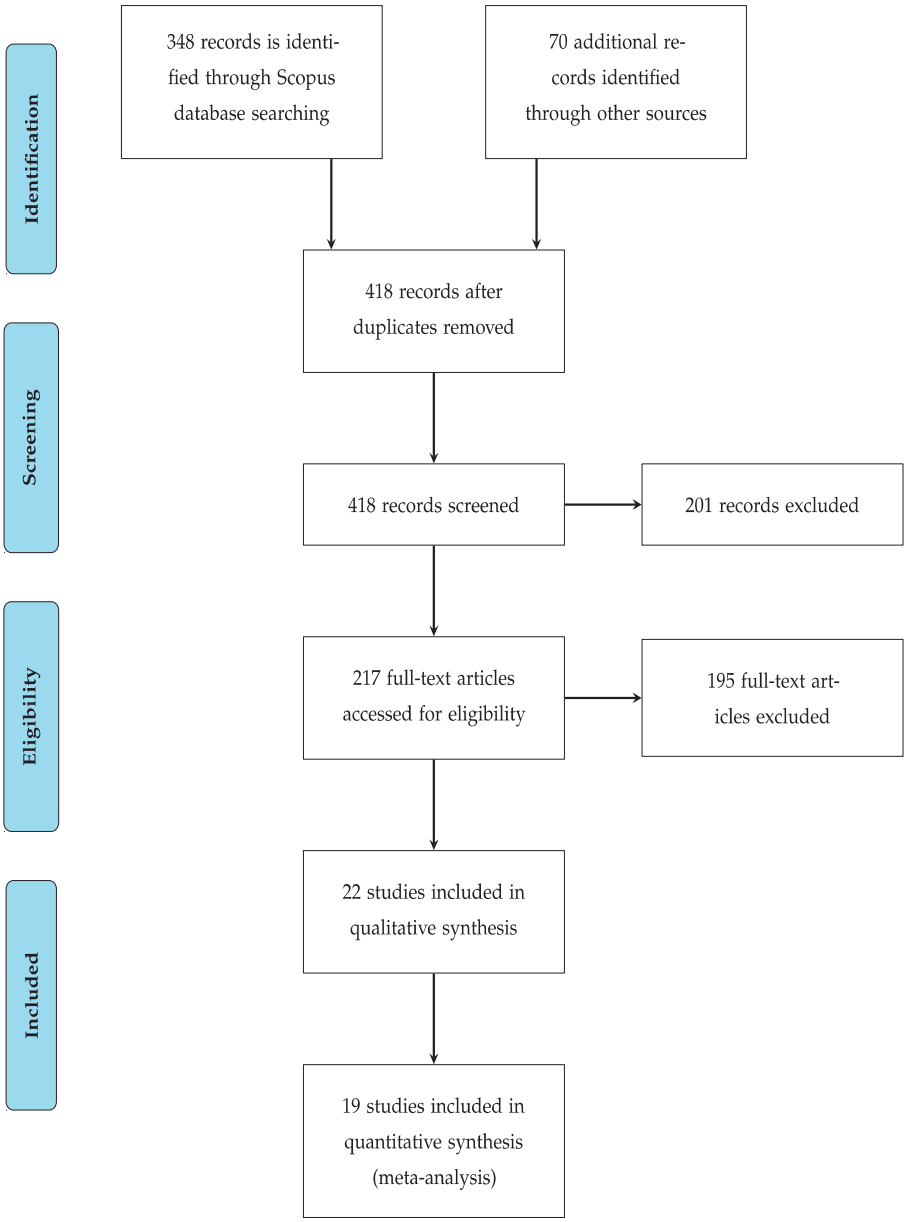

Figure 1.
Flow diagram

The dependent variable was the effect size: the Standardised Mean Difference (SMD) (Borenstein *et al.*, 2010). It measures the difference between the vulnerability of users in the experimental condition compared to the control group. Specifically, Cohen's *d* was used; this measure is the difference between the two means divided by the standard deviation for the data (Cohen, 2013).

The independent variables were categorised in three broad categories, as was done in the introduction:

(1) *context*: type of social engineering and pre-victimisation;

(2) *characteristics of the intervention*: modality intervention, priming, warning, focus, technical, format, tips, and intensity; and

(3) *characteristics of the evaluation study*: retention of knowledge, environment and randomised.

Based on the review above, we determined the following variables and coding categories.

Context

(1) Type of social engineering (categorical) measures via what device the attack reaches the target:
   - face-to-face (F2F);
   - email;
   - phone;
   - SMS; and
   - website.

(2) Pre-victimisation (categorical) measures whether subjects were victimised before receiving intervention:
   - No, all received training; and
   - Yes, Only victims were trained.

Characteristics of the intervention

(3) Modality intervention (categorical) measures what device was used to provide awareness:
   - orally, e.g. a spoken presentation;
   - static content, e.g. a pdf document; and
   - dynamic content based on the user's input, e.g. a game.

(4) Priming (categorical) measures whether an implicit warning against (cyber) crime/phishing was provided in the intervention:
   - No priming was used; and
   - Yes, the subjects were primed.

(5) Warning (categorical) measures whether an explicit warning against (cyber) crime/phishing was provided in the intervention:
   - No warning was given;
   - Only a warning was given; and
   - A warning with additional training materials (or the warning was included in the materials).

(6) Focus (categorical) measures the focus of the intervention materials:
   - URL, e.g. how to recognise a legitimate and malicious URL;
   - Email, e.g. what are suspicious pointers in emails;
   - Both URL and Email;
   - Social engineering;

- General (cyber)crime; and
- Other.

(7) Technical (categorical) measures whether the intervention contains a technical element, e.g. websites are no longer reachable by the subject:

- No technical elements were included; and
- Yes, there were technical elements.

(8) Format (categorical) measures the format that was used to deliver the awareness to the receiver:

- Text, the awareness was in textual form only, no use of graphics or illustrations;
- Text + graphics, the text contained graphics or illustrations to enhance understandability;
- Comic, a comic was used to provide the awareness (i.e. the use of graphics combined with speech balloons);
- Game, a game was used to provide the awareness; including a quiz or Q&A element where the subject receives feedback on the in-game performance to improve learning; and
- Other, another way was used, e.g. bookmarks.

(9) Tips (categorical) measures weather during the awareness tips or recommendations were provided to reduce victimisation:

- No tips were given;
- Only tips were given; and
- Tips with additional training materials (or the tips were included in the materials).

(10) Intensity (categorical) measures the amount of effort for the subject to complete one instance of the intervention:

- Low, e.g. information with tips;
- Medium, e.g. reading materials; and
- High, e.g. a lecture, a game which includes Q&A or training.

Methodological aspects of the study

(11) Retention of knowledge (continuous) measures the time between intervention and mock attack in hours:

- direct (post) test; and
- one-hour delay.

(12) Environment (categorical) measures in what environment the mock attack was performed:

- Lab is considered when research subjects are asked to come to a location that is designated and pre-setup for data collection for the experiment.
- Real-life is considered when a person participates as a research subject in a natural environment.

(13) Aware (categorical) measures to what extent people are aware of participating in an experiment:

- People are not aware that they participate in an experiment.
- People are aware that they are participating in an experiment; however, they are unaware of the goal of the study.
- People are aware both their participation in the experiment and objective of the study.

(14) Randomisation (categorical) measures whether the subjects were randomised across conditions: Based on the Maryland Scientific Method Scale (Farrington *et al.*, 2002), we distinguish between randomised experiments, quasi-experiment (non-randomised) and other methodologies.

- Not applicable. There was no randomisation of subjects over conditions used, or there was no mention of randomisation (e.g. studies that used a pretest-posttest design are an example of the former);
- Quasi-experiment; and
- Experiment: subjects were randomised among conditions (A quasi-experiment is similar to an experiment but does not randomise the units of analysis).

Readers

Two researchers independently coded the data items in all interventions. The first researcher (i.e. the first author) holds a PhD in computer science and has a background in both psychology and computer science. The second researcher (i.e. the second author) is a professor of cybersecurity and business continuity and has a background in psychology. An inter-rater reliability analysis using the Kappa statistic was performed to determine the consistency among researchers.

Procedures

All the interventions were coded twice, meaning that the resulting dataset consists of consensual results. After coding items, the inter-rater agreement was calculated. The scores of both readers were compared to generate the final data set. If both readers identified the same category for a given variable, there was a consensus. However, when there was a difference in the codes, the readers discussed the different views and reached a conclusion (the majority of differences related to one coder accidentally marking the wrong category).

Inter-rater agreement

The researchers' inter-rater reliability was: N = 605, $\kappa = 0.864$, 95% CI = [0.825, 0.903], $p = 0.000$. The amount of agreement indicates that we can reject the hypothesis that they are making their determinations randomly. The results indicate there is an almost perfect agreement between the two researchers (Landis and Koch, 1977).

Data analysis

Standardised Mean Difference (SMD) and confidence intervals were used as a measure of effect for the interventions. The pooled SMDs were computed using the *metafor* package in R (Viechtbauer, 2010). Twelve subgroup analyses were performed, one for each of the independent variables. Heterogeneity was assessed using the I^2 statistic. The I^2 values of 25, 50 and 75% indicate low, moderate and high heterogeneity, respectively. All analyses were conducted using a random-effects model. In our case, the studies were gathered from the published literature, and the true effect size varies from study to study. We assume that the

studies have enough in common that it makes sense to synthesise the information. Therefore, the random-effects model is the appropriate model to use (Borenstein *et al.*, 2010).

Results

Study characteristics

In total, there are 19 studies included in the analysis (Aburrous *et al.*, 2010; Arachchilage *et al.*, 2016; Bullée *et al.*, 2016; Bullée *et al.*, 2015; Caputo *et al.*, 2014; Jansson and von Solms, 2013; Jensen *et al.*, 2017; Junger *et al.*, 2017; Krol *et al.*, 2012; Kumaraguru *et al.*, 2009, 2008; Kunz, 2016; Lastdrager *et al.*, 2017; Lin *et al.*, 2011; Parsons *et al.*, 2015; Sheng *et al.*, 2010; Siadati *et al.*, 2017; Stockhardt *et al.*, 2016; Yang *et al.*, 2012). The studies were published between 2007 and 2017, for an overview of the studies per year, refer to Figure 2. Eleven studies provided a single effect size, two studies that provided two effect sizes, two studies that provided three effect sizes and four studies that provided four effect sizes.

Results of individual studies

In total, 19 studies were included in the analysis, with $N = 23{,}146$ subjects, having $k = 37$ observations (i.e. effect sizes/SMDs). Interventions to counter social engineering were associated with medium, significant reduction of victimisation with an effect size of 0.54 (95% CI = [0.359, 0.719], $I^2 = 89.31\%$, 37 studies). For an overview of effect sizes, refer to Figure 3 and for an overview of study characteristics, refer to Table A1. An SMD of 0.54 is usually considered to be a medium effect size, meaning that an average intervention reduces the likelihood of a user falling for a social engineering attack by 0.54 of the standard deviation of the specific outcome measure (Cohen, 2013).

Risk of bias across studies

Publication bias is the tendency that positive findings are more likely to be published. Negative or non-significant findings are often hard to publish, even in the case of a replication study (Kraemer and Andrews, 1982). The fail-safe N is a measure to indicate the number of zero effect studies required to impact the meta-analytical findings (Orwin, 1983). To reduce the found effect by half (i.e. SMD = 0.27) requires at least 37 zero effect studies to counter the effect of the present study.

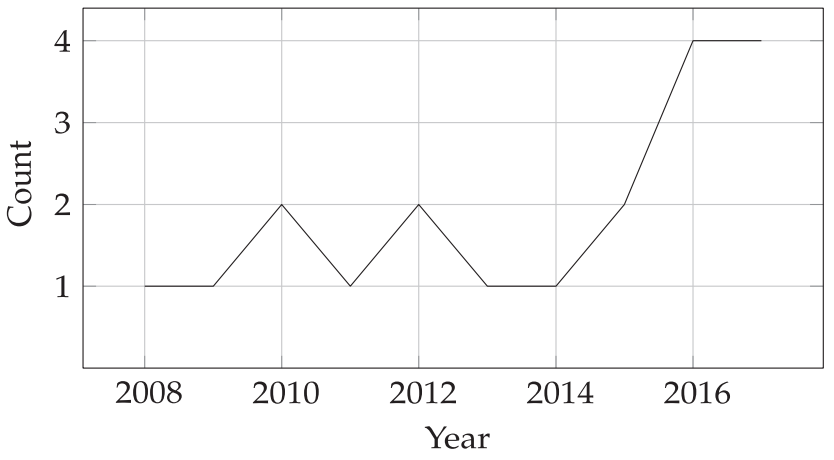

Figure 2.
Number of studies per year included in the analysis

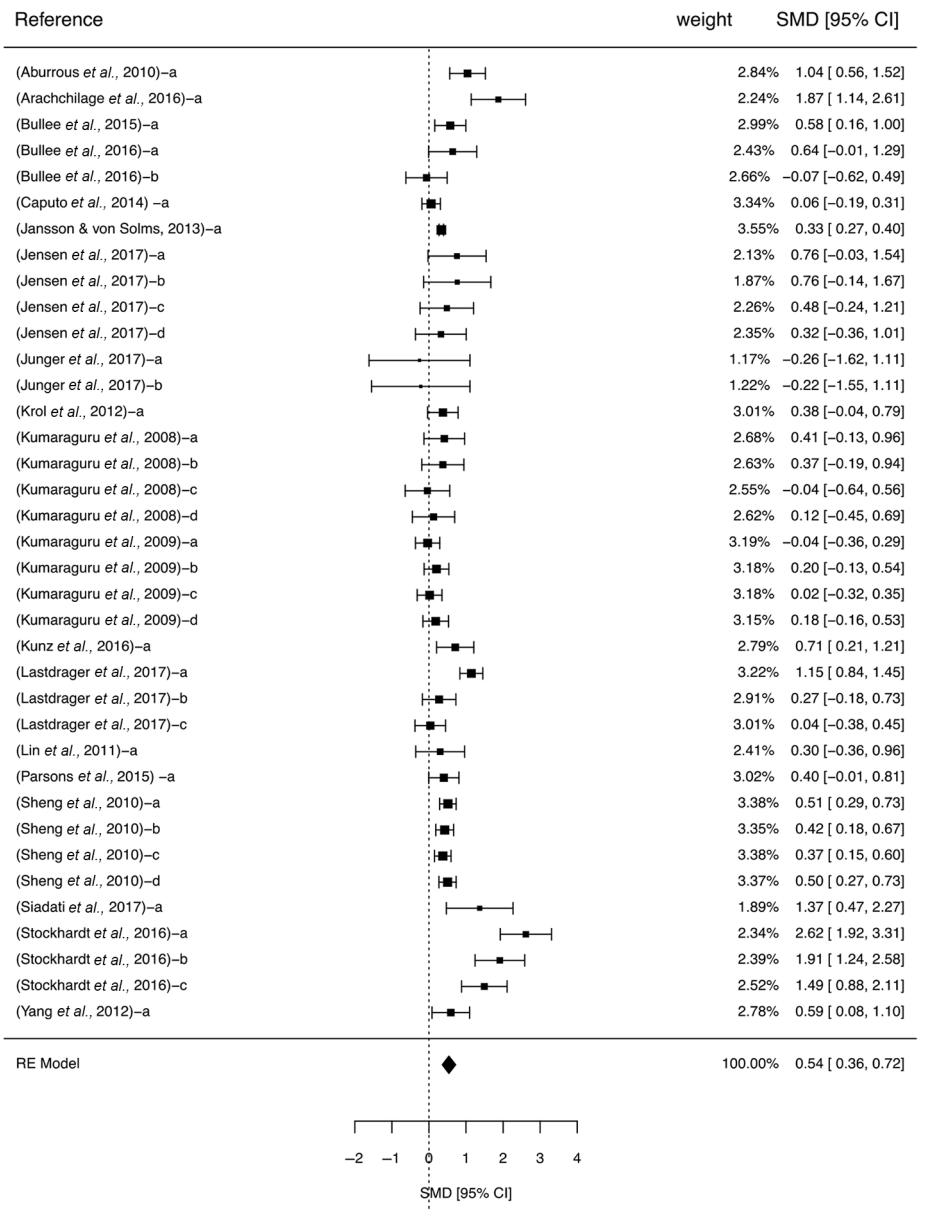

Figure 3.
Forest plot of both individual and pooled effect sizes in the meta-analysis, its weight and 95% CI; the dashed line indicates d = 0

Subgroup analysis

For each of the three groups of variables (i.e. context, characteristics of the intervention and characteristics of the testing method), results of the subgroup analysis are presented. For an overview of the results of the subgroup analysis, refer to Table 1.

Characteristic	SMD	95% CI	*n*	I^2 (%)	*p*
All	0.539	[0.359, 0.719]	37	89.31	–
Context					
Type of social engineering					0.002
Face-to-face	0.296	[−0.321, 0.913]	3	28.19	
Email	0.347	[0.224, 0.470]	23	70.40	
Telephone	0.267	[−0.423, 0.958]	2	61.68	
SMS	1.370	[0.470, 2.270]	1	0.00	
Website	1.250	[0.678, 1.823]	8	87.45	
Pre-victimised					0.002
No	0.703	[0.474, 0.933]	27	37.54	
Yes	0.182	[0.057, 0.307]	10	85.33	
Characteristics of the intervention					
Modality intervention					0.039
Spoken	0.995	[−0.119, 2.110]	4	96.18	
Static document	0.361	[0.235, 0.488]	27	63.78	
Dynamic content	0.941	[0.416, 1.465]	6	90.39	
Priming					0.001
No	0.383	[0.249, 0.517]	28	76.06	
Yes	1.013	[0.471, 1.555]	9	86.98	
Warning					0.848
No	0.533	[−0.160, 1.227]	3	61.52	
Warning only	0.343	[0.283, 0.402]	8	0.00	
Warn + train	0.580	[0.341, 0.819]	26	89.44	
Focus					0.008
URL	1.186	[0.676, 1.697]	9	88.79	
Email	0.336	[0.274, 0.397]	2	0.00	
URL + email	0.282	[0.116, 0.447]	16	67.23	
Social engineering	0.340	[−0.062, 0.743]	4	38.60	
Cybercrime	0.598	[0.055, 1.142]	3	0.00	
Other	0.521	[0.329, 0.712]	3	0.00	
Technical					0.675
No	0.547	[0.360, 0.734]	36		
Yes	0.334	[0.272, 0.396]	1		
Format					0.057
Text	0.651	[0.413, 0.889]	11	52.14	
Comic	0.232	[0.016, 0.447]	4	55.47	
Game	0.941	[0.416, 1.465]	6	90.39	
Text + comic	0.207	[0.078, 0.336]	11	10.01	
Other	0.858	[−0.044, 1.760]	5	94.68	
Tips					0.837
No	0.340	[0.279, 0.400]	5	0.00	
Tip only	0.429	[−0.037, 0.896]	5	59.73	
Tip + train	0.578	[0.350, 0.806]	27	88.81	
Intensity					0.038
Low	0.408	[0.157, 0.659]	9	71.23	
Medium	0.343	[0.190, 0.497]	18	52.28	
High	0.967	[0.463, 1.472]	10	93.80	

(*continued*)

Table 1. Average effect size d for sub-groupings of characteristics of 37 studies in the meta-analysis

Characteristic	SMD	95% CI	n	I^2 (%)	p
Characteristics of the testing method					
Retention of knowledge	$\beta = -0.0005$		37		0.047
Environment					0.007
Lab	0.810	[0.460, 1.160]	16	91.32	
Real-life	0.325	[0.168, 0.483]	21	71.72	
Aware					0.004
No	0.227	[0.109, 0.346]	14	35.49	
Half	0.402	[0.016, 0.789]	7	69.88	
Yes	0.870	[0.547, 1.193]	16	89.91	
Randomisation					0.913
n/a	0.478	[0.134, 0.822]	8	78.51	
Quasi	0.543	[0.168, 0.918]	5	42.42	
Randomised	0.576	[0.332, 0.819]	24	89.59	

Table 1.

Context

Type of social engineering

Interventions that were tested via SMS or website were associated with a very large effect in reducing victimisation (SMD = 1.370 and 1.250, respectively). Moreover, those that were tested via email, face-to-face and telephone were associated with a small effect (SMD = 0.347, 0.296 and 0.267, respectively). There was a statistically significant effect of the modality used to test the intervention on the reduction of victimisation for the five categories ($F(4, 32) = 5.534$, $p = 0.002$).

Pre-victimisation

The interventions performed regardless of someone was prior victimised were associated with medium to large effects in reducing victimisation (SMD = 0.703). In contrast, interventions performed after subjects were already victimised (i.e. embedded) were associated with small effects in reducing victimisation (SMD = 0.182). The effect of embedded interventions was statistically significantly smaller than the effect of non-embedded interventions ($Q(1) = 9.384$; $p = 0.002$).

Characteristics of the intervention

Modality intervention

Interventions that were presented verbally and those that used dynamic content were associated with a large effect on reducing victimisation (SMD = 0.995 and 0.941, respectively). Those that used static text were associated with a small to medium effect (SMD = 0.361). There was a statistically significant effect of the modality the awareness was presented on the reduction of victimisation for the three categories ($F(2, 34) = 3.570$, $p = 0.039$).

Priming

Interventions that use priming were associated with large effects in reducing victimisation (SMD = 1.013). However, those that do not use priming were associated with a small effect in reducing victimisation (SMD = 0.383). These differences were statistically significant ($Q(1) = 10.423$, $p = 0.001$).

Warning
Warnings, alone or in combination, did not have an impact on the effect SMD ($F(2, 34) = 0.166$, $p = 0.848$).

Focus
Interventions that focused on the URL were associated with a relatively large effect (SMD = 1.186), whereas those that focused on the topic of cybercrime were associated with a medium effect (SMD = 0.598). The interventions that focused on social engineering and those that focused on the content of an email were associated with a small to medium effect on reducing victimisation (SMD = 0.340 and 0.336, respectively). Those that focused on both the URL and email were associated with a small effect (SMD = 0.282). Finally, interventions that focused on another topic were associated with a medium effect (SMD = 0.521). There was a statistically significant effect of the focus of the awareness on the reduction of victimisation for the six categories ($F(5, 31) = 3.840$, $p = 0.008$).

Technical
Since only one study was identified that used built-in technical countermeasures, no subgroup analysis was performed.

Format
Interventions that used a combination of text and a comic were associated with large effects in reducing victimisation (SMD = 0.941), whereas those that used written statement (i.e. text) were associated with medium effects in reducing victimisation (SMD = 0.651). Interventions that used a comic and those that used a game were associated with small effects in reducing victimisation (SMD = 0.232 and 0.207, respectively). Finally, interventions that used another format were associated with a large effect (SMD = 0.858). There was – on the verge of – a statistically significant effect of the presented intervention format on the reduction of victimisation for the five categories ($F(4,32) = 2.568$, $p = 0.057$).

Tips
Interventions that used tips and tips in combination with training were associated with medium effects (SMD = 0.429 and 0.578, respectively). Those that did not provide tips were associated with small-to-medium effects in reducing victimisation (SMD = 0.340). There was not a statistically significant effect of the use of tips on the reduction of victimisation for the three categories ($F(2, 34) = 0.179$, $p = 0.837$).

Intensity
Interventions with high intensity were associated with a large to a very large effect (SMD = 0.967), whereas those with a low and medium intensity were associated with a small-to-medium effect in reducing victimisation (SMD = 0.408 and 0.343, respectively). This was a statistically significant effect of intensity ($F(2, 34) = 3.598$, $p = 0.038$).

Methodological aspects of the study

Retention of knowledge
The time between providing and testing the awareness is negatively associated with a very small significant reduction in victimisation ($p = 0.047$). The effect size decreases (SMD = −0.0005) for every hour that is between the intervention and the measure. Data had a minimum value of 0 (i.e. direct post-test) and a maximum value of 2160 (i.e. 90 days).

Environment real-life or lab

The interventions performed in a laboratory were associated with large effects in reducing victimisation (SMD = 0.810), whereas studies in a field setting were associated with small effect in reducing victimisation (SMD = 0.325). The effects of laboratory studies were statistically significant larger than the effects of field-tested interventions ($Q(1) = 7.188$; $p = 0.007$).

Aware

The degree to which subjects were aware of the fact that they participated in a study as a research subject mattered for the impact of the interventions. When subjects were aware of their participation, the intervention had a relatively high impact on reducing victimisation (SMD = 0.870); when they were somewhat aware, the intervention had a medium effect size (SMD = 0.402). Finally, when participants were unaware, the intervention was associated with a relatively low effect (SMD = 0.227). There was a significant effect of awareness on the reduction of victimisation for the three categories ($F(2,34) = 5.064$, $p = 0.012$).

Randomised

Whether studies randomised their subjects, mattered not for the impact of the intervention: Randomised studies had an effect size of 0.576, Quasi-randomised studies had an effect size of 0.543 and studies that did not use any randomisation had an effect size of 0.478. There was no statistically significant effect in the use of randomisation on the reduction of victimisation for the three categories ($F(2, 34) = 0.092$, $p = 0.913$).

Discussion

The present meta-analysis showed that interventions to reduce victimisation from social engineering were associated with a statistically significant medium effect. The standardised mean difference was 0.54, which is considered to be a medium effect size (Cohen, 2013). The results contrast to the somewhat negative views of authors that reviewed the impact of interventions in this field (Bada *et al.*, 2015; Ceesay *et al.*, 2018).

Importantly, when looking at the context, the intervention characteristics and the study characteristics, relatively large differences were found. In our sample, the distribution of categories in subgroups was skewed. Some categories had many observations, whereas others had only a few.

In the context of computer science and information/cyber security, there are systematic reviews performed (Yli-Huumo *et al.*, 2016; Agyepong *et al.*, 2019; Chockalingam *et al.*, 2017). A meta-analysis, however, builds on top of a systematic review and uses statistic methods to quantitatively pool and summarise the results of these studies (Akhter *et al.*, 2019). However, we believe that this is the first study that compared the impact of interventions of different type of social engineering in a systematic way. Comparing our findings with others is therefore not possible.

The effects of the intervention differed among the various types of social engineering. Interventions that focussed on social engineering attacks via SMS and websites were associated with higher effects compared to interventions that focussed social engineering attacks via email, telephone or F2F.

Providing interventions to only those who were previously victimised was associated with lower effects compared to providing interventions to the entire sample. This finding is not in line with what of Kumaraguru *et al.* (2007) reported. They stated that an embedded intervention is more effective than a non-embedded intervention.

One explanation could be that there is a moderating variable, possibly environment; all studies in the pre-victimisation group were tested in the lab, whereas those that were not in the pre-victimisation group were in both the field and lab.

Interventions that use a static document were associated with smaller effect size compared to interventions using dynamic content. Those that use spoken content were associated with the highest effects. From the perspective of the recipient, interventions that use spoken content can be perceived as more personal, tailored to their needs and specific situation. There is research supporting this finding (Stockhardt *et al.*, 2016). The importance of modality on memory has been demonstrated in an experiment where the subjects had to remember and recall a list of words or auditory representations. The results showed that the auditory representations had both a significant better recall and recall order of the presented stimuli (Drewnowski and Murdock, 1980). Glenberg (1984) showed that this auditory modality effect is also present in long-term memory; this means that auditory stimuli were better remembered in the long term compared to their visual counterparts (Glenberg, 1984). Furthermore, the positive effect of learning by gaming is confirmed by learning science (Clark and Mayer, 2016). Ideally, games should include examples, the execution of tasks with questions and answers, feedback and explanations when an answer is not correct. Also, users should have a form of control in game-based training. Proper training should allow its user to determine his or her pace (Clark and Mayer, 2016).

Interventions that used priming were associated with higher effects compared to those that did not use priming. These findings are in line with (Kenrick *et al.*, 2005; Dolan *et al.*, 2010).

The use of a warning had only a small impact on the effect size and is in line with the findings of Krol *et al.* (2012). They argued that people tend not to pay attention to warnings. When manipulating one condition (e.g. by providing a warning or intervention), it is important to have a manipulation check to be sure that the subjects noticed it. None of the studies that used warnings did also use a manipulation check. Therefore, the important question is did the people notice the warning? It has often been found that people do ignore (or do not notice) warnings (Krol *et al.*, 2012; Wu *et al.*, 2006). The cause of "goal hierarchy" could explain these findings; people tend to focus on their primary goal (Junger *et al.*, 2017; Krol *et al.*, 2012). Furthermore, give less attention to secondary goals, such as working securely. Usually, people are working or using their computer for a specific task. A security warning is often an unwelcome interruption of this task, which is why people tend to ignore them (Krol *et al.*, 2012).

Interventions that focus on identifying fraudulent URLs were associated with the highest effect compared to other elements. One explanation is that URLs are critical in the online realm. Everywhere one wants to go online, requires a URL. Being able to distinguish legitimate from fraudulent URLs is relatively easy and prevents one from much trouble. The instructions to determine a fraudulent URL are straight forward (e.g. be alerted when only numbers are used) compared to how to recognise a fraudulent email (e.g. is there a pressing tone of voice used) (Sheng *et al.*, 2010, 2007).

The effects of interventions differed among the different formats that were used. Interventions that consisted of a game were associated with the highest effect, compared to interventions that used a comic or a comic in combination with text, which were associated with the lowest effects. The format that uses only text had a larger effect compared to those who used a comic and text + comic. This finding is not in line with the findings of Kumaraguru *et al.* (2008), where the comic performed better. One explanation could be the limited number of observations (i.e. four effect sizes) that only used a comic as an intervention.

Providing a single tip or a combination of tips with training was equally effective. Here the same line of reasoning applies as for warnings; users may not be interested in tips that disrupt their activities. Alternatively, it could be that the tips were too abstract or complicated, and therefore, an ambiguous memory cue was created. Furthermore, it is also possible that once the subject is in the mock attack, there is no proper recollection of the cue to memory and therefore fails to recollect the tip.

Interventions that had a high intensity were associated with higher effects compared to those with medium and low intensity. The high-intensity interventions were, for example, a lecture or a game which includes questions and answers. Whereas, the interventions with a low or medium intensity more often contained more textual information. It could be that in the high-intensity interventions, there was more attention to the content of the materials. Attention is important in the encoding part of creating a memory. When attention is divided, encoding will be weaker, and future attempts to recollect memories are less successful (Smith and Kosslyn, 2008, p. 202).

The effect of an intervention decays over time. The existence of the memory decay was illustrated in multiple contexts; three letter nonsense syllables (Ebbinghaus, 1913), aeronautical knowledge (Casner *et al.*, 2006) and cardiopulmonary resuscitation (CPR) (Broomfield, 1996) and is now extended to cybercrime. Together with the findings, this constitutes the argument that awareness campaigns should be performed regularly (Bullée *et al.*, 2016).

Interventions that were tested in a laboratory setting were associated with higher effects compared with interventions tested in the field. Several studies, in economics, suggest that laboratory experiments, with their controlled conditions, are somewhat better able to produce high effect size, than field studies (Camerer *et al.*, 2016; Levitt and List, 2007). This does not mean that interventions in the lab cannot work in the field. However, this should be demonstrated (Shadish *et al.*, 2002).

There was no difference in effect size for studies that used a quasi-experiment, a randomised experiment or studies that did not randomise their sample. These findings contradict the findings of Weisburd *et al.* (2001) and Welsh *et al.* (2011). One possible explanation could relate to the scope of the study. The current study focused on cybercrime and social engineering specifically, whereas other studies focus on criminal behaviour in general.

In sum, a variety of effect sizes was found for interventions to decrease social engineering vulnerability. The analysis of intervention building blocks revealed that high-intensity interventions were more effective than low-intensity interventions. Furthermore, narrowly focused interventions were more effective than broadly focused interventions. In conclusion, interventions can counter social engineering attack; however, some interventions are more effective than others.

Limitations

The present study has several limitations:

- First, it could be that there are interventions that did not surface in the database search. However, the fail-safe N suggests that 37 non-effect studies are required to reduce the found effect by half.
- Second, multivariate analysis was not possible due to the limited sample size in the individual subgroups.
- Third, one should be careful when generalising the results. The vast majority of the studies was performed using subjects from WEIRD (Western, Educated,

Industrialised, Rich and Democratic) countries. These only constitute 10–15% of the world population (Henrich *et al.*, 2010). Therefore, generalising the results should be limited to WEIRD countries.

Future research

We present two suggestions for future research. First, the findings in this meta-analysis show what elements in interventions are effective and which features are less effective. We, therefore, suggest developing (and testing) interventions that combine the elements that are most effective and measure their effectiveness.

Second, the studies included are mainly from western countries and cultures. A suggestion for future research is to perform these studies in other countries and other cultures as well.

Implications for practitioners

We suggest using the findings of the present study to review the current awareness campaign organisations. For example, if the current anti-phishing training focusses on the content of the email, shift the focus towards the URL part of the email (SMD mail = 0.336, SMD URL = 1.186).

Alternatively, a new intervention can be developed, based on this study's findings. The different "characteristics of the intervention" subgroups can be used as a building block for this. For example, the use of priming proved to be more effective than not using priming; therefore, priming could be an element to include. Moreover, focussing the content to a specific subject is more effective than trying to cover a broad subject. Having a narrow focussed topic should be considered.

This analysis includes 19 studies and provides an initial overview of interventions. We believe that, although this is a great start, the number of relevant included studies is limited. To further increase the understanding of social engineering interventions, more data and more tested interventions are needed. By building a shared knowledge base contributes to both helping organisations and obtaining insight into social engineering interventions:

- Because the effectiveness of a new intervention cannot be easily predicted, organisations should perform mock attacks regularly to get insight into their susceptibility to social engineering attacks.
- When organisations are vulnerable, make interventions an element of these mock attacks and track their progress.
- Systematically perform the mock attacks and carefully record the procedure, conditions and findings. Aim to record characteristics of the training and context (e.g. the employee was in a shared office during the attack) that could explain the result.
- There is a lack of information on the effectiveness of interventions. Contribute the results of tested interventions anonymously to the knowledge base.

Conclusion

This meta-analysis found that people could benefit from interventions that counter victimisation caused by social engineering attacks. The evidence is gathered from interventions that were tested both in lab and field settings. As social engineering can have

devastating results for both the employee and the organisation, we must learn from past research and improve our interventions.

References

Abraham, S. and Chengalur-Smith, I. (2010), "An overview of social engineering malware: TRENDS, tactics, and implications", *Technology in Society*, Vol. 32 No. 3, pp. 183-196, available at: https://doi.org/10.1016/j.techsoc.2010.07.001

Aburrous, M., Hossain, M.A., Dahal, K. and Thabtah, F. (2010), "Experimental case studies for investigating e-banking phishing techniques and attack strategies", *Cognitive Computation*, Vol. 2 No. 3, pp. 242-253, available at: https://doi.org/10.1007/s12559-010-9042-7

Acquisti, A., John, L.K. and Loewenstein, G. (2012), "The impact of relative standards on the propensity to disclose", *Journal of Marketing Research*, Vol. 49 No. 2, pp. 160-174, available at: https://doi.org/10.1509/jmr.09.0215

Agyepong, E., Cherdantseva, Y., Reinecke, P. and Burnap, P. (2019), "Challenges and performance metrics for security operations center analysts: a systematic review", *Journal of Cyber Security Technology*, Vol. 0 No. 0, pp. 1-28, available at: https://doi.org/10.1080/23742917.2019.1698178

Akhter, S., Pauyo, T. and Khan, M. (2019), "What is the difference between a systematic review and a meta-analysis? ", in Musahl, V., Karlsson, J., Hirschmann, M. T., Ayeni, O. R., Marx, R. G., Koh, J. L., and Nakamura, N. (Eds.), *Basic Methods Handbook for Clinical Orthopaedic Research: A Practical Guide and Case Based Research Approach*, pp. 331-342, available at: https://doi.org/10.1007/978-3-662-58254-1_37

Alnajim, A. and Munro, M. (2009), "An anti-phishing approach that uses training intervention for phishing websites detection", *ITNG 2009 – 6th International Conference on Information Technology: New Generations*, pp. 405-410 available at: https://doi.org/10.1109/ITNG.2009.109

Arachchilage, N.A.G. and Cole, M. (2011), "Design a mobile game for home computer users to prevent from 'phishing attacks'", *Information society (i-society), 2011 International Conference on*, *IEEE*, pp. 485-489.

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, available at: https://doi.org/10.1016/j.chb.2016.02.065

Argo, J.J. and Main, K.J. (2004), "Meta-analyses of the effectiveness of warning labels", *Journal of Public Policy and Marketing*, Vol. 23 No. 2, pp. 193-208, available at: https://doi.org/10.1509/jppm.23.2.193.51400

Bada, M. Sasse, A.M. and Nurse, J.R.C. (2015), "Cyber security awareness campaigns: why do they fail to change behaviour?", Retrieved from www.cs.ox.ac.uk/files/7194/csss2015_bada_et_al.pdf

Blakeborough, L. and Correia, S. (2017), "The scale and nature of fraud: a review of the evidence", Retrieved from www.gov.uk/government/publications/the-scale-and-nature-of-fraud-a-review-of-the-evidence

Borenstein, M., Hedges, L.V., Higgins, J.P.T. and Rothstein, H.R. (2010), "A basic introduction to fixed-effect and random-effects models for meta-analysis", *Research Synthesis Methods*, Vol. 1 No. 2, pp. 97-111, available at: https://doi.org/10.1002/jrsm.12

Bosworth, S., Kabay, M.E. and Whyne, E. (2014), *Computer Security Handbook*, (6th ed.), Wiley, NY.

Broomfield, R. (1996), "A quasi-experimental research to investigate the retention of basic cardiopulmonary resuscitation skills and knowledge by qualified nurses following a course in professional development", *Journal of Advanced Nursing*, Vol. 23 No. 5, pp. 1016-1023, available at: https://doi.org/10.1111/j.1365-2648.1996.tb00084.x

Bullée, J.H., Montoya, L., Junger, M. and Hartel, P. (2017), "Spear phishing in organisations explained", *Information and Computer Security*, Vol. 25 No. 5, pp. 593-613, available at: https://doi.org/10.1108/ICS-03-2017-0009

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, available at: https://doi.org/10.3233/978-1-61499-617-0-107

Bullée, J.H., Montoya, L., Pieters, W., Junger, M. and Hartel, P.H. (2015), "The persuasion and security awareness experiment: reducing the success of social engineering attacks", *Journal of Experimental Criminology*, Vol. 11 No. 1, pp. 97-115, available at: https://doi.org/10.1007/s11292-014-9222-7

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

Cameron, C.D., Brown-Iannuzzi, J.L. and Payne, B.K. (2012), "Sequential priming measures of implicit social cognition: a meta-analysis of associations with behavior and explicit attitudes", *Personality and Social Psychology Review*, Vol. 16 No. 4, pp. 330-350, available at: https://doi.org/10.1177/1088868312440047

Camp, L.J., Grobler, M., Jang-Jaccard, J., Probst, C., Renaud, K. and Watters, P. (2019), "Measuring human resilience in the face of the global epidemiology of cyber attacks", *Proceedings of the 52nd Hawaii International Conference on System Sciences*.doi: 10.24251/HICSS.2019.574.

Campbell, D.T. and Stanley, J.C. (1963), *Experimental and Quasi-Experimental Designs for Research*, Houghton, Mifflin; Company, Boston, MA.

Canova, G., Volkamer, M., Bergmann, C. and Reinheimer, B. (2015), "NoPhish app evaluation: Lab and retention study", *NDSS workshop on usable security 2015*.

Caputo, D.D., Pfleeger, S.L., Freeman, J.D. and Johnson, M.E. (2014), "Going spear phishing: exploring embedded training and awareness", *IEEE Security and Privacy*, Vol. 12 No. 1, pp. 28-38, available at: https://doi.org/10.1109/MSP.2013.106

Casner, S.M., Heraldez, D. and Jones, K.M. (2006), "Retention of aeronautical knowledge", *International Journal of Applied Aviation Studies*, Vol. 6 No. 1, pp. 71-98.

Ceesay, E.N., Myers, K. and Watters, P.A. (2018), "Human-centered strategies for cyber-physical systems security", *ICST Transactions on Security and Safety*, Vol. 4 No. 14, pp. e5, available at: https://doi.org/10.4108/eai.15-5-2018.154773

Chan, M., Woon, I. and Kankanhalli, A. (2005), "Perceptions of information security in the workplace: Linking information security climate to compliant behavior", *Journal of Information Privacy and Security*, Vol. 1 No. 3, pp. 18-41, available at: https://doi.org/10.1080/15536548.2005.10855772

Chockalingam, S., Pieters, W., Teixeira, A., van Gelder, P. and, (2017), "Bayesian network models in cyber security: a systematic review", in Lipmaa, H., Mitrokotsa, A. and Matulevičius, R. (Eds.), *Secure It Systems*, Springer International Publishing, Cham, pp. 105-122.

Christin, N., Egelman, S., Vidas, T. and Grossklags, J. (2011), "It's all about the benjamins: an empirical study on incentivizing users to ignore security advice", *International Conference on Financial Cryptography and Data Security*, pp. 16-30 available at: https://doi.org/10.1007/978-3-642-27576-0_2

Clark, R.C. and Mayer, R.E. (2016), *E-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning*, John Wiley and Sons.

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

Davinson, N. and Sillence, E. (2010), "It won't happen to me: promoting secure behaviour among internet users", *Computers in Human Behavior*, Vol. 26 No. 6, pp. 1739-1747, available at: https://doi.org/10.1016/j.chb.2010.06.023

Dodge, R.C., Carver, C. and Ferguson, A.J. (2007), "Phishing for user security awareness", *Computers and Security*, Vol. 26 No. 1, pp. 73-80, available at: https://doi.org/10.1016/j.cose.2006.10.009

Dolan, P. Hallsworth, M. Halpern, D. King, D. and Vlaev, I. (2010), "MINDSPACE: influencing behaviour for public policy".

Downs, J.S., Holbrook, M.B. and Cranor, L.F. (2006), "Decision strategies and susceptibility to phishing", *Proceedings of The Second Symposium on Usable Privacy and Security*, pp. 79-90, available at: https://doi.org/10.1145/1143120.1143131

Drewnowski, A. and Murdock, B. (1980), "The role of auditory features in memory span for words", *Journal of Experimental Psychology: Human Learning and Memory*, Vol. 6 No. 3, pp. 319-332.

Ebbinghaus, H. (1913), *Memory: A Contribution to Experimental Psychology*, Teachers College, Columbia University.

Egelman, S. and Schechter, S. (2013), "The importance of being earnest [in security warnings]", *International Conference on Financial Cryptography and Data Security*, pp. 52-59, available at: https://doi.org/10.1007/978-3-642-39884-1_5

Egelman, S., Cranor, L.F. and Hong, J. (2008), "You've been warned: an empirical study of the effectiveness of web browser phishing warnings", *Proceedings of the Sigchi Conference on Human Factors in Computing Systems*, pp. 1065-1074, available at: https://doi.org/10.1145/1357054.1357219

Farrington, D.P., Gottfredson, D.C., Sherman, L.W. and Welsh, B.C. (2002), "The Maryland scientific methods scale", *Evidence-Based Crime Prevention*, pp. 13-21.

Feder, L., Jolin, A. and Feyerherm, W. (2000), "Lessons from two randomized experiments in criminal justice settings", *Crime and Delinquency*, Vol. 46 No. 3, pp. 380-400, available at: https://doi.org/10.1177/0011128700046003007

Furnell, S. (2007), "Phishing: Can we spot the signs?", *Computer Fraud and Security*, Vol. 2007 No. 3, pp. 10-15, available at: https://doi.org/10.1016/S1361-3723(07)70035-0

Glaspie, H.W. and Karwowski, W. (2018), "Human factors in information security culture: a literature review", in Nicholson, D. (Ed.), *Advances in Human Factors in Cybersecurity*, Springer International Publishing, pp. 269-280.

Glenberg, A. (1984), "A retrieval account of the long-term modality effect", *Journal of Experimental Psychology: Learning Memory and Cognition*, Vol. 10 No. 1, pp. 16-31, available at: https://doi.org/10.1037/0278-7393.10.1.16

Grazioli, S. (2004), "Where did they go wrong? An analysis of the failure of knowledgeable internet consumers to detect deception over the internet", *Group Decision and Negotiation*, Vol. 13 No. 2, pp. 149-172, available at: https://doi.org/10.1023/B:GRUP.0000021839.04093.5d

Grazioli, S. and Wang, A. (2001), "Looking without seeing: understanding unsophisticated consumers' success and failure to detect internet deception", *ICIS 2001 Proceedings*, p. 23.

Hadnagy, C. and Wilson, P. (2010), *Social Engineering: The Art of Human Hacking*, Wiley, New York, NY.

Happ, C., Melzer, A. and Steffgen, G. (2016), "Trick with treat – reciprocity increases the willingness to communicate personal data", *Computers in Human Behavior*, Vol. 61, pp. 372-377, available at: https://doi.org/10.1016/j.chb.2016.03.026

Heartfield, R. and Loukas, G. (2018), "Detecting semantic social engineering attacks with the weakest link: implementation and empirical evaluation of a human-as-a-security-sensor framework", *Computers and Security*, Vol. 76, pp. 101-127, available at: https://doi.org/10.1016/j.cose.2018.02.020

Henrich, J., Heine, S.J. and Norenzayan, A. (2010), "The weirdest people in the world?", *Behavioral and Brain Sciences*, Vol. 33 No. 2-3, pp. 61-83, available at: https://doi.org/10.1017/S0140525X0999152X

Henson, B., Reyns, B.W. and Fisher, B.S. (2016), "Cybercrime victimization", in *The Wiley Handbook on the Psychology of Violence*, pp. 553-570, available at: https://doi.org/10.1002/9781118303092.ch28

Herzberg, A. and Jbara, A. (2008), "Security and identification indicators for browsers against spoofing and phishing attacks", *ACM Transactions on Internet Technology (TOIT)*, Vol. 8 No. 4, p. 16, available at: https://doi.org/10.1145/1391949.1391950

Herzberg, A. and Margulies, R. (2013), "Forcing Johnny to login safely", *Journal of Computer Security*, Vol. 21 No. 3, pp. 393-424, available at: https://doi.org/10.3233/JCS-130467

Internet Crime Complaint Center (2018), "2017 Internet crime report", Retrieved from FederalBureauofInvestigationwebsitehttps://pdf.ic3.gov/2017_IC3Report.pdf

Jansson, K. and von Solms, R. (2013), "Phishing for phishing awareness", *Behaviour and Information Technology*, Vol. 32 No. 6, pp. 584-593, available at: https://doi.org/10.1080/0144929X.2011.632650

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, available at: https://doi.org/10.1080/07421222.2017.1334499

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

Kearney, W. and Kruger, H. (2014), "Considering the influence of human trust in practical social engineering exercises", *2014 Information Security for South Africa*, pp. 1-6, available at: https://doi.org/10.1109/ISSA.2014.6950509

Kenrick, D.T., Neuberg, S.L. and Cialdini, R.B. (2005), *Social Psychology: Unraveling the Mystery*, Pearson Education: New Zealand.

Kirlappos, I. and Sasse, M.A. (2012), "Security education against phishing: a modest proposal for a major rethink", *IEEE Security and Privacy Magazine*, Vol. 10 No. 2, pp. 24-32, available at: https://doi.org/10.1109/MSP.2011.179

Klahr, R., Shah, J., Sheriffs, P., Rossington, T., Pestell, G., Button, M. and Wang, V. (2017), "Cyber security breaches survey", *A Survey Detailing Business Action or Cyber Security and the Costs and Impacts of Cyber Breaches and Attacks. United Kingdom: Department for Culture, Media; Sport*, Vol. 2017.

Kraemer, H.C. and Andrews, G. (1982), "A nonparametric technique for meta-analysis effect size calculation", *Psychological Bulletin*, Vol. 91 No. 2, p. 404, available at: https://doi.org/10.1037/0033-2909.91.2.404

Krol, K., Moroz, M. and Sasse, M.A. (2012), "Don't work. Can't work? why it's time to rethink security warnings", *Risk and Security of Internet and Systems (Crisis), 2012 7th International Conference on*, pp. 1-8, available at: https://doi.org/10.1109/CRISIS.2012.6378951

Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L.F. and Hong, J. (2008), "Lessons from a real world evaluation of anti-phishing training", *ECrime Researchers Summit, 2008*, 1-12. *IEEE*.

Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L.F. and Hong, J. (2010), "Teaching johnny not to fall for phish", *ACM Transactions on Internet Technology*, Vol. 10 No. 2, pp. 1-31, available at: https://doi.org/10.1145/1754393.1754396

Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L.F., Hong, J. and Nunge, E. (2007), "Protecting people from phishing: the design and evaluation of an embedded training email system", *Conference on Human Factors in Computing Systems – Proceedings*, pp. 905-914, available at: https://doi.org/10.1145/1240624.1240760

Kumaraguru, P., Rhee, Y., Sheng, S., Hasan, S., Acquisti, A., Cranor, L.F. and Hong, J. (2007), "Getting users to pay attention to anti-phishing education: evaluation of retention and transfer", *ACM International Conference Proceeding Series*, 269, pp. 70-81, available at: https://doi.org/10.1145/1299015.1299022

Kumaraguru, P., Cranshaw, J., Acquisti, A., Cranor, L., Hong, J., Blair, M.A. and Pham, T. (2009), "School of phish: a real-world evaluation of anti-phishing training", *SOUPS 2009 – Proceedings*

of the 5th Symposium on Usable Privacy and Security, available at: https://doi.org/10.1145/1572532.1572536

Kunz, A., Volkamer, M., Stockhardt, S., Palberg, S., Lottermann, T. and Piegert, E. (2016), "Nophish: evaluation of a web application that teaches people being aware of phishing attacks", in Mayr, M., Heinrich C. and Pinzger, A (Ed.), *Informatik 2016*, (pp. 509-518). Bonn: Gesellschaft für Informatik e.V.

Landis, J.R. and Koch, G.G. (1977), "The measurement of observer agreement for categorical data", *Biometrics*, Vol. 33 No. 1, pp. 159-174, available at: https://doi.org/10.2307/2529310

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

Levitt, S.D. and List, J.A. (2007), "What do laboratory experiments measuring social preferences reveal about the real world?", *Journal of Economic Perspectives*, Vol. 21 No. 2, pp. 153-174, available at: https://doi.org/10.1257/jep.21.2.153

Lin, E., Greenberg, S., Trotter, E., Ma, D. and Aycock, J. (2011), "Does domain highlighting help people identify phishing sites?", *Proceedings of the Sigchi Conference on Human Factors in Computing Systems*, pp. 2075-2084, available at: https://doi.org/10.1145/1978942.1979244

Marinos, L. and Sfakianakis, A. (2012), *ENISA Threat Landscape*, ENISA.

Mayhorn, C.B. and Nyeste, P.G. (2012), "Training users to counteract phishing", *Work*, Vol. 41 No. Supplement 1, pp. 3549-3552, available at: https://doi.org/10.3233/WOR-2012-1054-3549

Monahan, T. and Fisher, J.A. (2010), "Benefits of 'observer effects': lessons from the field", *Qualitative Research*, Vol. 10 No. 3, pp. 357-376, available at: https://doi.org/10.1177/1468794110362874

Orwin, R.G. (1983), "A fail-safe *n* for effect size in meta-analysis", *Journal of Educational Statistics*, Vol. 8 No. 2, pp. 157-159, available at: https://doi.org/10.2307/1164923

Parsons, K., McCormac, A., Pattinson, M., Butavicius, M. and Jerram, C. (2015), "The design of phishing studies: challenges for researchers", *Computers and Security*, Vol. 52, pp. 194-206, available at: https://doi.org/10.1016/j.cose.2015.02.008

Parsons, K., Calic, D., Pattinson, M., Butavicius, M., McCormac, A. and Zwaans, T. (2017), "The human aspects of information security questionnaire (hais-q): two further validation studies", *Computers and Security*, Vol. 66, pp. 40-51, available at: https://doi.org/10.1016/j.cose.2017.01.004

Reep-van den Bergh, C.M.M. and Junger, M. (2018), "Victims of cybercrime in europe: a review of victim surveys", *Crime Science*, Vol. 7 No. 1, p. 5, available at: https://doi.org/10.1186/s40163-018-0079-3

Rouse, M. (2006), "Definition social engineering", Retrieved from www.searchsecurity.techtarget.com/definition/social-engineering

Schmidt, R.A. and Bjork, R.A. (1992), "New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training", *Psychological Science*, Vol. 3 No. 4, pp. 207-218, available at: https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

Schneier, B. (2000a), "Crypto-gram, October 15, 2000", Retrieved from www.schneier.com/crypto-gram/archives/2000/1015.html

Schneier, B. (2000b), *Secrets and Lies: Digital Security in a Networked World*, (1st ed.). New York, NY: John Wiley and Sons, Inc.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin.

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L.F. and Downs, J.S. (2010), "Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions", *Conference on Human Factors in Computing Systems– Proceedings*, 1, pp. 373-382, available at: https://doi.org/10.1145/1753326.1753383

Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L.F., Hong, J. and Nunge, E. (2007), "Anti-phishing Phil: the design and evaluation of a game that teaches people not to fall for phish", *Proceedings of The 3rd Symposium on Usable Privacy and Security*, pp. 88-99, available at: https://doi.org/10.1145/1280680.1280692

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, available at: https://doi.org/10.1016/j.cose.2016.09.009

Siedler, T. and Sonnenberg, B. (2010), *Experiments, Surveys and the Use of Representative Samples as Reference Data (No. 146,*), German Council for Social; Economic Data (RatSWD).

Smith, E.E. and Kosslyn, S.M. (2008), *Cognitive Psychology: Mind and Brain*, Pearson Prentice Hall.

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

Sundar, S.S., Kang, H., Wu, M., Go, E. and Zhang, B. (2013), "Unlocking the privacy paradox: do cognitive heuristics hold the key?", *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pp. 811-816, available at: https://doi.org/10.1145/2468356.2468501

The Federal Bureau of Investigation (2015), "Business email compromise", Retrieved from www.ic3.gov/media/2015/150827-1.aspx

Verizon Risk Team (2018), Retrieved from www.verizon.com/about/sites/default/files/2018-Verizon-Annual-Report.pdf

Viechtbauer, W. (2010), "Conducting meta-analyses in r with the metafor package", *Journal of Statistical Software*, Vol. 36 No. 3, pp. 1-48, available at: https://doi.org/10.18637/jss.v036.i03

Weisburd, D., Lum, C.M. and Petrosino, A. (2001), "Does research design affect study outcomes in criminal justice? ", *The ANNALS of the American Academy of Political and Social Science*, Vol. 578 No. 1, pp. 50-70, available at: https://doi.org/10.1177/000271620157800104

Welsh, B.C., Peel, M.E., Farrington, D.P., Elffers, H. and Braga, A.A. (2011), "Research design influence on study outcomes in crime and justice: a partial replication with public area surveillance", *Journal of Experimental Criminology*, Vol. 7 No. 2, pp. 183-198, available at: https://doi.org/10.1007/s11292-010-9117-1

Winkler, I.S., Dealy, B. (1995), "Information security technology?. . .don'T rely on it: a case study in social engineering", *Proceedings of the 5th Conference on Usenix Unix Security Symposium – Volume*, *5*, 1-1. Berkeley, CA: USENIX Association.

Wogalter, M.S., Laughery, K.R., Sr,. and Mayhorn, C.B. (2012), "Warnings and hazard communications", *Handbook of Human Factors and Ergonomics*, pp. 868-894, available at: https://doi.org/10.1002/9781118131350.ch29

Wolff, J. (2016), "Perverse effects in defense of computer systems: when more is less", *Journal of Management Information Systems*, Vol. 33 No. 2, pp. 597-620, available at: https://doi.org/10.1080/07421222.2016.1205934

Wu, M., Miller, R.C. and Garfinkel, S.L. (2006), "Do security toolbars actually prevent phishing attacks?", *Proceedings of the Sigchi Conference on Human Factors in Computing Systems*, pp. 601-610, availabel at: https://doi.org/10.1145/1124772.1124863

Xiao, B. and Benbasat, I. (2015), "Designing warning messages for detecting biased online product recommendations: an empirical investigation", *Information Systems Research*, Vol. 26 No. 4, pp. 793-811, available at: https://doi.org/10.1287/isre.2015.0592

Yang, C.C., Tseng, S.S., Lee, T.J., Weng, J.F. and Chen, K. (2012), "Building an anti-phishing game to enhance network security literacy learning", *2012 IEEE 12th International Conference on Advanced Learning Technologies*, pp. 121-123, available at: https://doi.org/10.1109/ICALT.2012.174

Yli-Huumo, J., Ko, D., Choi, S., Park, S. and Smolander, K. (2016), "Where is current research on blockchain technology? – A systematic review", *Plos One*, Vol. 11 No. 10, pp. 1-27, available at: https://doi.org/10.1371/journal.pone.0163477

Zhang, B. and Xu, H. (2016), "Privacy nudges for mobile applications: effects on the creepiness emotion and privacy attitudes", *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, pp. 1676-1690, available at: https://doi.org/10.1145/2818048.2820073

Zhang, B., Wu, M., Kang, H., Go, E. and Sundar, S.S. (2014), "Effects of security warnings and instant gratification cues on attitudes toward mobile websites", *Proceedings of the Sigchi Conference on Human Factors in Computing Systems*, pp. 111-114, available at: https://doi.org/10.1145/2556288.2557347

Corresponding author
Jan-Willem Bullee can be contacted at: j.h.bullee@gmail.com

Appendix

Table A1. Summary of interventions

Reference – measurement	*N*	SMD	Type	PreVic	Modality	Priming	warning[1]	Focus[2–4]	format[5]	Tips[1]	Intensity	Delay	Environ	Aware	Random
Aburrous *et al.* (2010) – 1	22	1.689	PC	No	Document	Yes	No	URL	Text	Yes	Medium	0	Lab	Half	Quasi
Arachchilage *et al.* (2016) – 1	20	1.780	PC	No	Dynamic	No	Yes + *t*	URL	Text + c	Yes + *t*	High	0	Lab	Half	N/A
Bullée *et al.* (2015) – 1	118	0.528	F2F	No	Document	No	Yes	SE	Text	Yes + *t*	Medium	168	Real	No	Yes
Bullée *et al.* (2016) – 1	64	0.641	Web	No	Document	No	Yes	SE	Comic	Yes	Medium	168	Real	No	Yes
Bullée *et al.* (2016) – 2	63	−0.065	Web	No	Document	No	Yes	SE	Comic	Yes	Medium	336	Real	No	Yes
Caputo *et al.* (2014) – 1	329	0.543	Mail	Yes	Document	No	Yes + *t*	URL + e	Comic	Yes + *t*	Medium	2160	Real	No	Yes
Jansson *et al.* (2013) – 1	17504	0.344	Mail	Yes	Document	No	Yes	Mail	Comic	No	Low	168	Real	No	N/A
Jensen *et al.* (2017) – 1	73	0.757	Mail	No	Document	No	Yes + *t*	CC	Text	Yes + *t*	Medium	240	Real	Half	Yes
Jensen *et al.* (2017) – 2	60	0.764	Mail	No	Document	No	Yes + *t*	CC	Game	Yes + *t*	Medium	240	Real	Half	Yes
Jensen *et al.* (2017) – 3	67	0.641	Mail	No	Document	No	Yes + *t*	URL + e	Text	Yes + *t*	Medium	240	Real	Half	Yes
Jensen *et al.* (2017) – 4	66	0.322	Mail	No	Document	No	Yes + *t*	URL + e	Game	Yes + *t*	Medium	240	Real	Half	Yes
Junger *et al.* (2017) – 1	184	0.111	F2F	No	Document	Yes	No	CC	Text	No	Low	0	Lab	Yes	Quasi
Junger *et al.* (2017) – 2	190	0.096	F2F	No	Document	No	Yes	SE	Text	Yes	Low	0	Lab	Yes	Quasi
Krol *et al.* (2012) – 1	120	0.376	PC	No	Document	No	Yes	Other	Text	No	Low	0	Lab	Half	Yes
Kumaraguru *et al.* (2008) – 1	152	0.414	Mail	Yes	Document	No	Yes + *t*	URL + e	Game	Yes + *t*	Medium	48	Real	No	N/A
Kumaraguru *et al.* (2008) – 2	149	0.374	Mail	Yes	Document	No	Yes + *t*	URL + e	Game	Yes + *t*	Medium	48	Real	No	N/A
Kumaraguru *et al.* (2008) – 3	152	−0.042	Mail	Yes	Document	No	Yes + *t*	URL + e	Game	Yes + *t*	Medium	168	Real	No	N/A
Kumaraguru *et al.* (2008) – 4	149	0.122	Mail	Yes	Document	No	Yes + *t*	URL + e	Game	Yes + *t*	Medium	168	Real	No	N/A
Kumaraguru *et al.* (2009) – 1	261	−0.036	Mail	Yes	Document	No	Yes + *t*	URL + e	Game	Yes + *t*	Medium	168	Real	No	Yes
Kumaraguru *et al.* (2009) – 2	261	0.203	Mail	Yes	Document	No	Yes + *t*	URL + e	Game	Yes + *t*	Medium	672	Real	No	Yes
Kumaraguru *et al.* (2009) – 3	261	0.017	Mail	Yes	Document	No	Yes + *t*	URL + e	Game	Yes + *t*	Medium	168	Real	No	Yes
Kumaraguru *et al.* (2009) – 4	261	0.183	Mail	Yes	Document	No	Yes + *t*	URL + e	Game	Yes + *t*	Medium	672	Real	No	Yes
Kunz *et al.*, (2016) – 1	32	0.698	Mail	No	Dynamic	Yes	Yes + *t*	URL	Text + c	Yes + *t*	High	0	Lab	Half	N/A
Lastdrager *et al.* (2017) – 1	189	1.145	Mail	No	Spoken	No	Yes + *t*	URL + e	Other	Yes + *t*	High	0	Real	Yes	Yes
Lastdrager *et al.* (2017) – 2	81	0.273	Mail	No	Spoken	No	Yes + *t*	URL + e	Other	Yes + *t*	High	336	Real	Yes	Yes
Lastdrager *et al.* (2017) – 3	96	0.036	Mail	No	Spoken	No	Yes + *t*	URL + e	Other	Yes + *t*	High	672	Real	Yes	Yes
Lin, *et al.* (2011) – 1	100	0.154	PC	No	Document	Yes	No	URL	Other	Yes	Low	0	Lab	Yes	N/A
Parsons *et al.* (2015) – 1	117	0.347	Mail	No	Document	Yes	Yes	Mail	Text	No	Low	0	Lab	Half	Quasi
Sheng, *et al.* (2010) – 1	384	0.511	Mail	No	Document	No	Yes + *t*	Other	Text	Yes + *t*	Medium	0	Lab	Half	Yes
Sheng, *et al.* (2010) – 2	419	0.425	Mail	No	Dynamic	No	Yes + *t*	URL	Text + c	Yes + *t*	High	0	Lab	Half	Yes
Sheng, *et al.* (2010) – 3	417	0.502	Mail	No	Document	No	Yes + *t*	URL + e	Game	Yes + *t*	Medium	0	Lab	Half	Yes
Sheng, *et al.* (2010) – 4	435	0.502	Mail	No	Dynamic	No	Yes + *t*	URL + e	Text + c	Yes + *t*	High	0	Lab	Half	Yes
Siadati *et al.* (2017) – 1	52	1.506	Phone	No	Document	No	Yes	Other	Text	No	Low	0	Real	No	Yes
Stockhardt *et al.* (2016) – 1	30	2.615	PC	No	Spoken	Yes	Yes + *t*	URL	Other	Yes + *t*	High	0	Lab	Half	Yes
Stockhardt *et al.* (2016) – 2	25	1.913	PC	No	Dynamic	Yes	Yes + *t*	URL	Text + c	Yes + *t*	High	0	Lab	Half	Yes

(*continued*)

Table A1.

Reference – measurement	*N*	SMD	Type	PreVic	Modality	Priming	warning[1]	Focus[2–4]	format[5]	Tips[1]	Intensity	Delay	Environ	Aware	Random
Stockhardt *et al.* (2016) – 3	26	1.494	PC	No	Document	Yes	Yes + *t*	URL	Text	Yes + *t*	Medium	0	Lab	Half	Yes
Yang *et al.* (2012) – 1	62	0.591	PC	No	Dynamic	Yes	Yes + *t*	URL	Text + c	Yes + *t*	High	0	Lab	Yes	Quasi

Notes: For each effect size are the sample size, SME, context and characteristics shown [1]Yes + *t* = Yes and a training; [2]SE = Social Engineering; [3]CC = CyberCrime; [4]URL + e = URL + email; [5]Text + c = Text + Comic; PreVic = pre-victimised