A meta-analysis of field experiments on phishing susceptibility

Teodor Sommestad
C4ISR
Swedish Defence Research Agency
Linköping, Sweden
teodor.sommestad@foi.se

Henrik Karlzén
C4ISR
Swedish Defence Research Agency
Linköping, Sweden
henrik.karlzen@foi.se

***Abstract*—Phishing is a serious threat to any organization allowing their employees to use messaging systems and computers connected to the internet. Consequently, researchers have undertaken a large number of studies to identify the variables that determine this threat, i.e. variables that influence users' susceptibility to phishing emails. This paper presents a meta-analysis of the findings in 48 papers describing field experiments. The mean susceptibility rate to phishing emails across all studies and measurements was 21 percent. A majority (116 of 140) of the association tests reported, concerned variables related to the recipient. Most of these reported insignificant results. Both relative risks and association tests showed that technical warning systems, email personalization, training, and the use of established deceptive tactics influence the susceptibility rate. The type of scam as such also appears to be important, with some types of scams being orders of magnitude more successful than other types. Many of the results had limitations in control and sampling, which may explain unexpected and contradictory results.**

Keywords—phishing, fraud, social engineering, information security

I. INTRODUCTION

Phishing is the act of electronically and deceitfully contacting a person, with the aim of making the person electronically perform an act that is beneficial to the deceiver and harmful to the deceived. Phishing incidents typically occur by email. For example, an employee may receive an email from an address similar to that of the head of the IT-department, requesting the recipient to provide network credentials or execute malicious code. Doing so will put both the recipient and their organization at risk.

A. Phishing susceptibility

Phishing susceptibility, i.e. the probability that a recipient performs an action requested in a fraudulent message, is a widespread problem. According to the cyber breach data classified by Verizon, 32% of all breaches performed in 2018 involved phishing [1]. It is our belief that there are good reasons for the prevalence of phishing incidents. First, this type of attack is typically cheap to execute in comparison to other attacks that circumvent perimeter protection systems such as firewalls. Complex software is sometimes a part of the phishing operation, but it is not needed for the attackers to gain a foothold in the targeted system. Second, the attack procedure can be repeated multiple times and on multiple system users. This considerably increases the probability that some employee within an organization will be deceived. Third, as email is often used to make requests, a large portion of computer users are inclined to trust requests via email and actually try to perform the actions requested. Data collected in tests performed by the company Cofense indicate how susceptible people are in general. By sending 135 million synthetic phishing emails, Cofense managed to make employees visit websites, open attachments and perform other potentially risky behaviors in 12% of the cases during 2017 and 2018 [2].

Coping with the threat of phishing is difficult. The benefits of email and other electronic communication are considerable and, as [3] found by interviewing computer users, the strategy of not clicking on any links in emails received cannot be used indiscriminately, since such links may be perceived as necessary for work activities. However, there is a considerable variation in phishing susceptibility, depending on the recipient, situation, and phishing content (e.g. message). For instance, [4] reported that only 0.3% of targets were deceived by an email containing a credit card scam, whereas 37% were deceived by a scam about a course registration. The company Cofense reported that they only succeeded in 5.3% of phishing cases against employees in the energy industry, whereas the success rate was 15.0% in the health care industry [2]. Cofense also reported a considerable variation over time, indicating a situational factor. For example, the success rate of an online order email with an attachment was 4.4% in the first quarter of 2018, increasing to 18.4% in the second quarter. Such differences in susceptibility rates are remarkable. The aim of this paper is to summarize the existent knowledge concerning factors related to phishing susceptibility.

B. Research on phishing

The scholarly literature on phishing is dominated by studies on technical counter-measures [5]. This dominance is also visible in terms of available literature reviews. There are a number of reviews of technical counter-measures related to phishing, e.g. [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. However, no encompassing review of empirical studies of what influences computer users' susceptibility to phishing has been found.

Empirical studies on human susceptibility can be crudely classified into observational studies, laboratory experiments,

This research is sponsored by the Swedish Civil Contingencies Agency.

and field experiments. *Observational studies* often focus on the properties of phishing emails, based on historical emails classified as malicious (e.g. [16]), or the characteristics of phishing victims (e.g. [17]). However, it is difficult to observe all relevant factors in such studies. *Laboratory experiments* typically expose people to a situation resembling a real phishing case, e.g. by presenting them with a set of synthetic (albeit often based on historical real phishing) emails. These individuals are given different forms of training and are asked to identify which of the emails are malicious and which are benign. A number of experiments of this sort have been conducted, e.g. by [18]. Some laboratory experiments have used functional magnetic resonance imaging (fMRI) scanning to identify regions of the brain that are activated when distinguishing between legitimate and phishing websites [19]. These tests can provide detailed information about when people are able to distinguish phishing from other types of communication and variables that are of importance to phishing susceptibility. However, even if the experiment itself involves deception and participants are unaware of the researchers' aims, the situation in a laboratory is different from that in a normal workplace. This makes generalizations to office environments unreliable, especially when it comes to variables' absolute, rather than relative, influence on phishing susceptibility. The effect sizes obtained from *field experiments*, where synthetic phishing emails are sent to unaware users, are more ecologically valid (i.e. more naturalistic), and can therefore provide more information on effect sizes in practice. On the other hand, the natural environment of computer users is difficult to control, and matters such as spam filters, office hours, and incident handling performed by IT departments may present an obstacle to researchers. Indeed, field experiments have also reported varying susceptibility rates, as in the example of [4] mentioned above. Thus, while field experiments have the potential to produce ecologically valid results and representative susceptibility rates, they may be biased by uncontrolled variables.

C. Aims and scope

This paper synthesizes the results of published field experiments in order to obtain an average susceptibility rate, assessing how this rate is influenced by variables present in field settings, and describes issues related to experimental designs. Literature database searches yielded 48 papers describing susceptibility rates or statistical tests from field experiments on variables' relationship to susceptibility. The studies were reviewed to answer the following questions:

- What variables influence phishing susceptibility?
- What qualities and flaws are there in the research?

D. Paper outline

The rest of this paper is organized as follows: Section II presents a schematic overview of the types of variables that are believed to determine phishing susceptibility - the message, the recipient, and the situation. Section III describes the method used to identify and synthesize existing research. Section IV summarizes the susceptibility rates, significance tests, and relative risks reported in the literature. Section V relates these results to the theories presented in section II. Section VII discusses experimental designs and the qualities and flaws associated with them. Section VII discusses the reliability of the results and gives suggestions of what they might mean to researchers and practitioners.

II. Susceptibility to phishing messages

Cofense's assessments showed that there is a considerable variation in phishing susceptibility, e.g. with the susceptibility rate in one quarter more than four times the rate of another quarter [2]. As stated by the company regarding the relation to the variation in their data:

"Rates can depend on many factors, such as the number of simulations, by whom, and how often. As seen in the data on 'Invoice' phishes, timing can be a factor too, with users less alert during busy periods." [3, p. 18]

As this paper will show, there is a limited agreement concerning theoretical frameworks and models for predicting and explaining individuals' susceptibility to phishing. In addition, the subsequent sections will show that the results in phishing research often differ from predictions of general deception theories, and no established model exists for prediction of phishing susceptibility. To provide a basis for the rest of this paper, this section offers a brief overview of how susceptibility to phishing may be influenced by

1) the attributes of the message
2) the character of the recipient
3) the situation the recipient is in.

It is the authors' belief that variables associated with these three high-level factors can explain most of the variance in susceptibility. As this review will show, all variables studied in existent research can also be associated with these three factors, even though the focus of the research is unevenly distributed among them. However, it is important to note that this simple list of factors does not aspire to be an all-encompassing prediction model for phishing, but only serves as a way to classify variables that such a model may want to include. The text below aims to giving examples of how the three factors may influence phishing susceptibility, with some pointers to research on deception and persuasion in general, as well as concrete examples from phishing research.

A. The message

It is intuitive to think that the content and wording used in a phishing message influence the probability that a recipient will perform the requested action(s). Research on both deception and persuasion supports this notion, and research specifically on phishing confirms several of these more general theories. However, some results also contradict general theories on deception and persuasion. Examples of when established theories are confirmed and contradicted are given below.

The *self-presentational perspective on cues to deception* presented by DePaulo et al. [20] is one example of a general theory on deception. According to this theory, liars are less forthcoming (e.g., respond with less detail and seem to hold back), are more tense, less positive/pleasant, have fewer ordinary imperfections and unusual content in their stories, and have less compelling tales (e.g., have less engaging and fluent tales). Some evidence suggests that email recipients look for such cues and are more likely to be deceived if the cues are not

present (which may be the case if the liar has had more time to plan the lie). For example, [21] found that recipients were more susceptible to emails with content that was familiar and humorous (what [21] calls "liking"), i.e., were more pleasant. On the other hand, [21] obtained almost the same degree of susceptibility without a pleasant tone, but stressed the scarcity of a resource and requiring a swift action instead (9.1% vs. 10.3%).

A second example of relevant theory related to the message can be drawn from marketing research. Research on advertisements has shown that the use of an authority is an effective persuasion principle [22]. Accordingly, real phishing emails tend to use the subject field to refer to some authority [16]. On the other hand, pretending to be an authority in the body of emails was actually less effective than other influence techniques in the test described in [21]. This suggests that those sending real phishing emails (and relying heavily on authority references) either construct their emails in a sub-optimal way or that research has failed to refer to authority in an effective way.

A third example of research in other domains with relevance to phishing is the framework developed by Johnson et al. [23], relating to deception in a financial context. They stated that deception may involve the following techniques: masking (e.g. not disclosing an expense), dazzling (e.g. writing things in footnotes rather than the body), decoying (e.g. emphasizing irrelevant legal issues), repackaging (e.g. changing labels of economic entities), mimicking (e.g. creating fictitious transactions), and double play (improperly applying generally accepted accounting principles). When [24] adopted parts of this framework for a test on phishing, support for decoying, but not for dazzling and mimicking, was found.

B. The recipient

It is reasonable to expect that some people are easy to deceive while others are not. Theories and research on deception in general indeed demonstrate that there are relevant variables tied to the recipient, whether the variables are more stable (e.g. personality), or more malleable (e.g. knowledge). Such variables may reveal the existence of particularly vulnerable people in need of targeted efforts to reduce their susceptibility.

An example of a proposition related to stable variables can be found in *interpersonal deception theory* [25]. Among other things, this theory posits that the receiver's truth bias, i.e. their general inclination to believe that people are truthful and pleasant, is related to the accuracy of deception judgments. However, empirical tests on phishing do not provide clear and consistent results concerning this. For example, tests have reported insignificant correlations to personality traits such as agreeableness, pessimism [26], propensity to trust others [27], and helpfulness [28].

Malleable variables can also be related to phishing susceptibility. For example, laboratory experiments on deception in general have shown that a happy mood tends to make people more gullible than a sad or neutral mood [29]. The *interpersonal deception theory* [25] also covers malleable variables. In particular, it posits that recipients' skills to detect deceptive messages vary. However, the empirical evidence for this is ambiguous, and even if there are those who are more skilled, it is unclear how to find them. For instance, in general research on face-to-face deception, experts (e.g. law enforcement personnel and auditors) have been shown to be no more accurate at detecting lies than novices [30]. On the other hand, laboratory phishing experiments have reported positive correlations between detection accuracy and self-rated capacity in handling phishing (r=0.16), awareness of padlock icon in browsers (r=0.18), and intelligence (combining novel problem-solving and experience) (r=0.27) [31]. However, [28] found no significant correlation between susceptibility in field tests and susceptibility on test scenarios or computer experience. Similarly, [27] found no significant correlation between phishing susceptibility and competence, but instead found that a high level of internet usage increased susceptibility.

While the situation regarding skill is not yet clear, many phishing interventions are directed at reducing recipients' susceptibility using training or education. Training is sometimes embedded in tests so that those who are deceived by a phishing email receive training whereas others do not, e.g. as in [32] and [33]. The results related to training are predominantly positive and have shown a reduction in susceptibility after training (e.g. in [33], [34], [35]). However, there are studies with opposite results too. For instance, [32] obtained higher susceptibility with the last email in the training program than in the first email (before any training).

C. The situation

There are obvious and direct links between phishing susceptibility and some variables related to the situation the recipient is in when the deception (attempt) occurs. For instance, it may be that some recipients struggle to keep up with the inflow to their mailbox, e.g. due to receiving too many emails or being on vacation. Accordingly, [36] reported that a person experiencing a high email load is less likely to respond to phishing emails. On the other hand, [36] also indicated that high email loads may make users pay less attention to clues associated with phishing emails. Thus, a large amount of email in a user's inbox may make the user both more gullible and less susceptible to fraud because emails are left without action. In order not to underestimate phishing susceptibility, it makes sense to determine the likelihood of actions on legitimate emails. For instance, the tests by [33] were preceded by a legitimate informational email, in which only four out of ten recipients clicked the link.

Other situational variables of importance are those related to protection mechanisms (such as spam filters and warnings), since such mechanisms may prevent reception and processing of a phishing email. To complicate matters, it appears that some measures intended to protect can actually have an adverse effect on susceptibility. More specifically, [37] reported that many users clicked the link in their phishing email after a warning email was sent by the chief security officer asking employees to ignore it and the phishing email. The study speculated that this could be due to curiosity and unawareness about the risks associated with malicious links.

D. Interactions

Finally, it is reasonable to expect that interactions between variables associated with the abovementioned factors are of importance. Yet, few studies have explicitly addressed such

interactions between the message, the recipient, and the situation the recipient is in. However, it is possible to find examples of the importance of such interactions.

A number of studies have tested emails that are more or less adapted to the target population, i.e. an interaction between the message and the recipient. This is typically done by varying the message to make it appear as coming from a sender the recipient is more or less familiar with. In general, adaption to the target population makes phishing emails more successful. For example, [38] obtained a susceptibility rate of 72% when a phishing email that appeared to have been sent by a person the recipient knew, but only a susceptibility rate of 16% when the email instead seemed to have been sent by a stranger.

The interaction between the message and the situation has been largely overlooked in published research, with the exception of [39]. The qualitative analyses reported in [39] found that when users' work context aligned with the premise of the email, they found it more believable and focused on the more compelling parts of the email rather than the clues of deception. This was illustrated in tests using a phishing email about a missing voice call. Users who indeed had a missed call found the email more believable than those who knew that they had not missed a call. Another example was provided by [40], who sent emails to military cadets concerning grades at the end of the semester. They concluded that "timing was key in this experiment because of the subject of the phishing message." [39, p. 2]

The interaction between the recipient and the situation is another aspect that may be important, as some people respond differently than others in certain situations. There are a few studies explicitly addressing such interactions. For instance, there is research showing that some users are able to make more use out of warnings in browsers than others [31]. Thus, some counter-measures may only work for certain recipients.

III. Review Methodology

This section describes the search method used to identify relevant studies, the criteria used to determine which of the identified studies to include in the review, how data were extracted from the studies, and how these data were synthesized.

A. Search method

Literature was identified using systematic queries in the database Scopus and using keyword searches in Google Scholar. Searches in Scopus required the paper abstract, keyword or title to include both the word "phishing" and at least one of the words "test", "experiment", and "survey". Searches in Google Scholar used the phrase "phishing experiment". In March 2019, this returned 615 records from Scopus. Searches in Google Scholar yielded thousands of records, of which the first 200 were included.

The relevance of these records' contributions was determined based on their abstracts and titles. Among the records retrieved from Google Scholar, 73 were considered relevant and seemed to contain empirical data. The 615 records from Scopus yielded another 92 records that appeared to contain empirical studies of relevance. Together, this yielded 165 records (conference papers, articles, theses, reports).

B. Inclusion criteria

The 165 papers were downloaded in full text and their contributions were assessed in more detail. To be included in the review, a paper should:

1. Be written in English.
2. Concern phishing via email.
3. Describe results from a field experiment.
4. Report significance tests and/or susceptibility rates.

Of the 165 papers, 48 fulfilled all four criteria. The other papers were written in German (1 paper), unrelated to email phishing (3 papers), a literature review (1 paper), a research plan (2 papers), an empirical analysis based on expert judgment (1 paper), an observational study (1 paper), laboratory experiments (54 papers), questionnaires (52 papers), or both laboratory experiments and questionnaires (2 papers).

Quality was not among the inclusion criteria. Instead, a quality evaluation was used as a step in evaluating the second research questions (concerning qualities and flaws in the research).

C. Data extraction: susceptibility

Two types of empirical analyses were extracted from the papers: susceptibility rates under different conditions and statistical significance tests.

Susceptibility rates were not always straightforward to extract from the papers. For example, [41] and [32] reported sequences of tests in the same population, where only those susceptible to an email received special treatment before they received the next one. In those cases, only a subset of the data that was straightforward to interpret and compare with other data was extracted. For instance, only susceptibility rates of the first and last training emails were extracted from [32]. For each identified relevant data point, the following was extracted:

- The message (a short description of its content, the purported sender, and level of personalization).
- Treatment (e.g. a different version of the email or training prior to the experiment).
- Susceptibility rate (click link, provide password, open attachment, run executable, provide other information).

The tests for the association between susceptibility rates and other variables had been performed using different statistical methods, e.g. Pearson correlations, ANOVA, regression analyses, and T-tests. Consequently, the type of effect sizes reported in the studies varied considerably. As a result, only the direction and significance of the relationship were extracted. For each test, it was extracted whether the test gave a statistical significant difference and whether the difference increased or decreased phishing susceptibility. For the papers that used an inversely coded response variable (e.g. "deception detection"), the direction was reversed. Some papers reported mixed results for variables, e.g. using multiple tests where different variables were controlled for. Those cases were classified as mixed significance.

D. Data extraction: quality aspects

Based on the literature on experiment design, it is clear that there are a number of procedures researchers conducting and reporting experiments should adhere to, but sometimes do not. First, at least one hypothesis should be formulated. Second, the research design should be decided in advance of actually performing the experiment. Among other things, this entails , using power analysis to determine the likely needed sample size [42], pilot testing the design [43], and, sometimes, pre-registering the design [44]. Third, experiments need to randomize recruited participants into treatment and control groups, or otherwise use some sort of control for confounding variables [45] (e.g. using participants as their own control [43], [46]). Fourth, any experiment needs to be described in sufficient detail to allow for replication, e.g. making raw data available to other researchers, if possible [47]. Fifth, any ethical implications of the experiment for the stakeholders need to be considered, including approval by institutional review boards (IRBs) or others determining the shape of informed consent [48] and debriefing of participants (e.g. to eliminate participants' perception of harm [49]), and whether any ethical guidelines are used. Sixth, relevant laws should also be considered (e.g. laws on data protection [50], or on trademarks when imitating companies [51], [52]). As part of the data extraction, it was determined whether the papers included descriptions of these issues. More succinctly, this part of the data extraction tried to answer the following questions:

- Was there at least one explicit hypothesis?
- Was power analysis used to determine adequate sample size?
- Was pilot testing used?
- Was the experiment design pre-registered?
- Were participants randomized into control groups, or were confounding variables otherwise controlled?
- Was publication of raw data mentioned?
- Were ethical implications considered?
- Were legal issues considered?

In addition to these variables, it was assessed if the researchers ensured control of emails actually reaching the recipients, recipients reading the emails, and which individuals were deceived by the phishing attempt.

E. Data synthesis

The extracted data on phishing susceptibility were synthesized in three ways: as susceptibility rates under different circumstances, as tests of associations that are similar to each other, and as relative risks of different conditions for the susceptibility rates. Further, the differences between studies concerning experimental design were also investigated. More details are provided below.

1) Susceptibility rates

Susceptibility rates were obtained under a range of conditions that can be assumed to influence the results. As the number of possible configurations of the variables that differ between the field tests widely exceeds the number of susceptibility rates reported, an analysis based on a few high-level variables was considered most relevant. This analysis grouped the susceptibility rates into five dimensions (e.g. population and adaptation of the email). Not all papers were possible to classify according to the dimensions (e.g. the population was not described), and not all configurations had been studied. However, in the present analysis, summary statistics of reported susceptibility rates are reported for each configuration of these dimensions for those studies that could be classified.

When the number of effect sizes allowed, Medcalc was used to calculate sample-weighted susceptibility rates and 95% confidence intervals, using a random effects model. This summary statistic was calculated for heterogeneous populations. The susceptibility rates of all studies had an I^2 statistic of 99.2%, suggesting that almost all differences between studies' susceptibility rates were due to other factors than sampling error (chance). The I^2 statistic was above 90% for most of the classes used in this synthesis, suggesting that the susceptibility rates were contingent on more than the five variables mentioned above.

2) Signficance tests

The variables where association with susceptibility had been tested were often conceptually similar. For example, several papers tested if gender had a significant relationship to susceptibility. When variables were considered similar or identical, they were treated as the same variable in the analysis. For instance, variables related to gender were converted to represent the impact of the recipient or sender being female. Because of the great variety in statistical tests and reporting formats, a simple vote counting procedure was used to synthesize the findings of studies.

It is worth noting that not all cases were as straightforward as the gender example above. For instance, a number of self-reported scales measuring various types of security knowledge and security awareness were aggregated into one variable. A number of scales measuring threat perception, vulnerability perception, and risk perception were aggregated as another variable. In some cases, studies tested variants of the same concept, e.g. both competence and its opposite (incompetence) [27]. These were also synthesized, but with reversion of sign when needed.

3) Relative risks

Not all experimental designs were designed to control for apparent exogenous variables, leading to large differences between studies. However, experiments that vary a condition while keeping other conditions more or less constant are well suited for showing the relative importance of variables under those conditions. For example, some papers reported susceptibility to emails constructed differently within a population, such as generic versus population-adapted emails. The ratio of the susceptibility obtained with and without population-adaptation represents how population-adaption influences the relative risk associated with phishing.

4) Quality aspects

Variables associated with experimental design and quality were assessed as either being met, not met, or partially met.

These assessments were only aggregated to illustrate to what extent the included studies met the criteria.

IV. RESULTS

The study sample sizes varied from nine (in a pilot study) to 19,180. Median was 248, with a few studies having sample sizes of more than 10,000. The subsequent sections present synthesized data on phishing susceptibility, tested associations between phishing susceptibility and other variables, the relative risk assessed from paired measurements, and experimental designs.

A. Susceptibility rates

Table I describes mean susceptibility rates of all measurements reported in the studies. These susceptibility rates are grouped based on whether they are from a study performed in a university setting, involve a message purporting to be from someone the recipient should trust, if the recipient has received training (including previous experiments), to what extent the email was adapted to the recipient, and the action requested and measured by the researchers. Adaptation was classified as generic (G) when there was no adaptation at all, population (P) when the email was framed as relevant for a certain audience (e.g. students at a university), and individual (I) when it was framed as something sent only to the recipient (e.g. mentioning the recipient's name). Measured actions were classified as clicking a link to a website (L), providing a credential such as a password (C), providing other sensitive information (I), opening an attachment such as a pdf file (A), or executing code such as an EXE file (E). Some studies were difficult to classify according to this scheme. These studies are only included in the cases that do not discriminate between conditions (marked with a dash).

For all 145 measurements (the row with only dashes in the first five columns from the left), the mean weighted susceptibility rate was 21%, with a 95% confidence interval of 19%–24%. However, susceptibility rates varied considerably both within and between the studies performed under the same conditions. Susceptibility rates varied between 4% and 68% between the conditions and 95% confidence intervals could be as wide as 9%–69% for one specific condition. Some differences fell within the expected. For example, training seemed to be associated with lower susceptibility rate (21% vs. 28%) and it was more difficult to make users execute code than click on a link (2% vs 24%). However, other differences are surprising and counter-intuitive. For instance:

- Susceptibility was higher (36%) in the measurements where the sender was a stranger compared to measurements where the purported sender and receiver had a trust-relationship (20%).
- Recipients were almost as susceptible to providing their passwords (21%) as they were to clicking links (24%).
- Recipients were as susceptible to phishing emails targeting them specifically (e.g. saluting them by name) as they were to generic undirected phishing (17% for both).

TABLE I. SUSCEPTIBILITY UNDER DIFFERENT CONDITIONS.

University [a]	Sender trusted [a]	Training [a]	Adaptation [b]	Action [c]	Number of measurements	Number of studies	Total sample size	Susceptibility rate	Low (2.5 %)	High (97.5 %)
Y	Y	Y	I	L	4	2	30408	0.13	0.07	0.20
Y	Y	Y	I	C	9	3	30196	0.18	0.08	0.31
Y	Y	Y	P	L	16	6	6458	0.27	0.15	0.41
Y	Y	Y	P	C	16	9	13999	0.21	0.12	0.33
Y	Y	Y	P	I	4	3	11304	0.09	0.00	0.29
Y	Y	Y	P	E	2	1	1500	0.08	0.06	0.10
Y	Y	Y	G	L	4	2	1772	0.32	0.20	0.46
Y	Y	N	I	C	1	1	40	0.23	-	-
Y	Y	N	P	L	3	1	883	0.36	0.09	0.69
Y	Y	N	P	C	1	1	33	0.24	-	-
Y	N	Y	I	C	1	1	20	0.50	-	-
Y	N	Y	P	L	1	1	125	0.68	-	-
Y	N	Y	P	C	1	1	60	0.20	-	-
Y	N	Y	G	L	5	3	2028	0.42	0.27	0.58
N	Y	Y	I	L	1	1	158	0.27	-	-
N	Y	Y	I	E	1	1	158	0.09	-	-
N	Y	Y	G	L	5	4	4150	0.11	0.08	0.14
N	Y	Y	G	E	5	4	4150	0.05	0.04	0.06
N	N	Y	G	C	2	1	258	0.20	0.10	0.33
N	N	Y	G	I	1	1	129	0.04	-	-
N	N	Y	G	A	1	1	129	0.26	-	-
-	-	-	-	-	145	39	151810	0.21	0.19	0.24
Y	-	-	-	-	80	21	109434	0.23	0.19	0.27
N	-	-	-	-	53	12	14751	0.18	0.15	0.21
-	Y	-	-	-	77	24	106367	0.20	0.17	0.24
-	N	-	-	-	26	9	7555	0.36	0.29	0.44
-	-	Y	-	-	136	39	129308	0.21	0.18	0.24
-	-	N	-	-	7	3	1178	0.28	0.13	0.46
-	-	-	I	-	17	5	60980	0.17	0.13	0.22
-	-	-	P	-	71	25	46130	0.21	0.17	0.27
-	-	-	G	-	38	13	42006	0.17	0.14	0.21
-	-	-	-	L	72	24	59378	0.24	0.20	0.28
-	-	-	-	C	54	20	73684	0.21	0.17	0.27
-	-	-	-	I	7	5	12597	0.19	0.06	0.37
-	-	-	-	A	2	1	159	0.06	0.05	0.07
-	-	-	-	E	10	6	5992	0.02	0.02	0.03

[a] Y – Yes; N – No; [b] P – Population; I – Individual; G – Generic; [c] A – open Attachment; L – Click on link; C – provide credentials; E – Execute code; I – provide Information.

While these differences are surprising and counter-intuitive, it should be noted that they might very well be explained by differences in study designs and target populations. As already noted, the variation in measured susceptibility rates was high both within and between these groups of studies. Thus, variables of importance are clearly missing in this classification scheme, which may cause bias. Subsequent sections will present data less prone to such bias.

Another detail worth noting in Table I is the types of studies that researchers have performed. A majority of the emails in these tests were preceded by some kind of training or another measurement, were not spear phishing, measured users clicking links or providing passwords, purported to be from someone the recipient should trust, and were performed within universities.

B. Significance tests

The susceptibility rates described in Table I above are potentially biased due to differences in experimental conditions of the aggregated experiments. However, a number of studies reported statistical tests of variables associated with phishing susceptibility rates and were performed using more control, e.g. correlating characteristics of the recipients to susceptibility with one phishing email and within one population. The variables tested are shown in Table II, together with the number of studies with tests showing a statistically significant relationship (in a positive or in a negative direction), the number of studies with mixed results (e.g. depending on sample or test used), and the number of tests showing an insignificant relationship.

A large majority of the tests (116 of 140) concerned variables related to the recipient. The message (16 tests) and the situation (1 tests) were given less attention. Fourteen (14) tests dealt with interaction. The number of tests with insignificant or mixed results is considerable.

A full 67 of the 116 tests on recipient-related variables were insignificant, and six studies reported mixed significance. In fact, most variables related to the recipient were associated with contradictory reports. Only s*elf-reported behavior* and *cyber scholarship* had reports that were significant, and in one direction only (as they were measured only once). All the others were either associated with mixed or contradictory results.

Variables associated with the message showed clearer relationships to susceptibility, with mail richness being the only variable with mixed results. On the other hand, half of the tests associated with the message had only been tested in a single study, and 5 out of 16 tests gave insignificant results. This also applies to the well-established influence techniques of referring to reciprocity (e.g. asking for a returned favor), consistency (e.g. claiming that a behavior has been performed before), liking (e.g. being humorous), social proof (e.g. claim that others perform the action), and referring to an authority (e.g. purporting that a senior executive requires the action). Three of these were positively correlated to susceptibility, one was insignificant, and one was negatively correlated to susceptibility when the others, as well as gender, were controlled for [21].

Some interaction effects were found to be related to the recipient's gender, the sender, and the content. However, only one out of six tests resulted in clear results.

TABLE II. VARIABLES' ASSOCIATION WITH SUSCEPTABILITY.

Variable	Positive (increase)	Negative (decrease)	Mixed significance	Insignificant
The recipient (overall)	**18**	**23**	**5**	**67**
Propensity to trust	4			8
Self-reported security knowledge	1	5		5
Phishing training	1	6		3
Gender (female)	2		1	7
Education level		2	2	3
Risk perception	1			6
Computer experience		3		4
Age		2		4
Agreeableness	2			4
Openness	1			5
Susceptibility on paper scenarios	1			2
Conscientiousness	1			2
Neuroticism	1		1	1
Integrity				3
Intention to resist social engineering	1	1		
Risk behavior		1	1	
Computer self-efficacy		1		1
Use of other channels to confirm emails		1		1
Extraversion	1			1
Benevolence				2
Faculty status (not student)	1		1	
Self-reported susceptability	1			
Cyber scholarship		1		
Integrity (e.g. honesty)				1
Focus				1
Pessimism				1
Boredom proneness				1
Entertainment drive				1
The message (overall)	**9**	**2**	**0**	**5**
Email richness	2	1		
Urgency	2			
Loss threat	2			
Sent by relevant source				2
Sent by female				2
Refer to authority		1		
Refer to reciprocity	1			
Refer to consistency				1
Refer to liking	1			
Refer to social proof	1			
The situation (overall)		**1**		
Browser-warning		1		
Interactions (overall)	**6**	**0**	**1**	**7**
Personalization	5			2
National culture*Recipient variables				5
Sender female*Receiver female	1			
Recipient female*Scam used			1	

C. Relative risks

Many studies reported susceptibility rates obtained when one condition varied, while other conditions were kept constant. These susceptibility rates can be compared to obtain the relative risk (RR) associated with the varied condition. Six types of

conditions varied in this way were identified. These are listed in Table III along with the sample weighted mean relative risk and its 95% confidence interval.

As the table shows, recipients were more susceptible when a good scam (e.g. the storyline) was used (RR=5.9), when the email content was adapted to the recipient (RR=1.5), or an established deceptive tactic (e.g. referring to authority) was used (RR=2.6). Conversely, the risk was reduced when training was provided (RR=0.4), the user had a web browser extension that provides warnings when phishing messages are processed (RR=0.3), and when the email requests the recipient to take a more risky action (RR=0.9).

TABLE III. RELATIVE RISKS FROM PAIRED MEASUREMENTS.

Condition	Paired measurements	Sample size	Relative risk (RR)	Low (2.5 %)	9High (97.5 %)
A type of scam that works better in the target population is used [a]	39	34901	5.9	4.3	7.9
Adaptation is used to target the recipient better	13	46297	1.5	1.1	2.0
Phishing training has been provided	11	5387	0.4	0.1	1.3
An established deceptive tactic is used	8	8478	2.6	1.8	3.7
A more risky action is required by the recipient [b]	6	83932	0.9	0.3	2.5
A web browser extension with warnings is used	2	55	0.3	0.2	0.5

[a] These are arranged retrospectively so that the more successful scam is the riskier condition, i.e. so that the relative risk is always equal to or greater than one.

[b] Ranked in riskiness as: clicking a link < open attachment < provide sensitive information < provide password < execute code, with all increases in riskiness rating equally when compared to a less risky action.

D. Experimental designs

As may be gathered from Table IV, not all studies had clearly described hypotheses. In some cases, hypotheses were substituted by research questions that were either equivalent to hypotheses (see hypotheses in Table IV), or not (rated partially). In addition:

- Very few studies explicitly used power analysis to calculate the required sample size. Four studies did not calculate but used experience.
- It was relatively rare for studies to describe pilot testing of their research design (ten studies, plus two studies that were pilot studies in themselves).
- No study described pre-registering of the experiment design, e.g. with an academic journal.
- A mere eight papers clearly described having a control group. A further nine papers stated that they controlled some variables, but most did not mention this topic at all.
- Two papers mentioned making raw data available to other researchers (and two others proposing the construction of a database with shared research data).

Some papers mentioned checking email delivery, reading, and action. Only ten papers checked if emails reached their intended destination, rather than being stuck in spam filters, etc. Two papers consisted of two studies, and both only checked one of them. A mere five papers checked if emails were read (e.g. by read receipts). One paper that consisted of two studies checked in only one of them. Most papers checked which participants in the sample who were deceived by the scam (rather than merely checking how many were deceived). The checking was performed e.g. by registering the IP address or user name of the visitor to a website (or the unique link used to reach the website) linked in the email. One paper that consisted of two studies only checked one of them.

TABLE IV. EXPERIMENTAL DESIGN.

Criteria	Yes	Partially	No
Hypotheses	33	1	14
Pilot testing	10	2	36
Power analysis	3	4	41
Pre-registering	0	0	48
Control group	8	9	31
Raw data	2	0	46
Control of email delivery	10	2	36
Control of email being read	5	1	42
Control of which participants in the sample who were deceived by the scam	27	1	20

As shown in Table V, most studies at least mentioned or discussed the ethics of their research, albeit very briefly (seven cases, included as partially in the table). Slightly more than a third of the papers mentioned receiving approval by their institutional review board (IRB) (typically a university entity, but in some cases a government agency). In six further cases (plus one of the studies in one paper) where IRB approval was not mentioned, the papers at least mentioned some other kind of approval (e.g. by the management of the studied organization). Approval was also (separately) given through the participants' informed consent. However, exceedingly few papers described having acquired informed consent, with some others even explicitly mentioning that they did not seek informed consent. Of the few papers that did describe acquiring informed consent, one mentioned using a cover story, whereas another hid some details. One paper (included as partially in the table) contained one study that did have consent and one that did not.

TABLE V. ETHICAL AND LEGAL ISSUES.

Criteria	Yes	Partially	No
Discusses ethical issues	33	7	8
IRB approval	18	0	30
Other approval	6	1	41
Informed consent	3	1	44
Debriefing	17	0	31
Uses guidelines	8	0	40
Legal issues	6	0	42

Table V also shows that about a third of the papers described debriefing participants after their studies; eight papers

mentioned adhering to a guideline on ethical phishing research; and only six papers made any consideration of legal issues. Three papers met the experimental design quality criteria to a greater extent than the rest of the papers. Two of these were theses [34][53] (which typically contain more details) and the third stated that "the study was time consuming, taking almost one year to conduct" [54].

V. VARIABLES OF RELEVANCE

Section II of this paper introduced a model with four types of variables. These were tied to the message, the recipient, the situation, and interactions between the other three. This section discusses the relevance of these variables in light of the result of the meta-analysis.

A. The recipient

Overall, the recipient seemed to be less important than the researchers had anticipated. Reports concerning variables related to the recipient mainly involved statistical significance tests. In total, 113 reports on tested relationships were extracted. The majority of these relationships (64%) were either insignificant or had mixed results that depended on how the test was performed.

A fair share of the tests concerned variables related to personality, e.g. the big five personality traits. Most of these seemed to be insignificant for phishing susceptibility. This also applies to variables with a close conceptual relationship to deception, such as the propensity to trust others and agreeableness. Other tests concerned self-rated knowledge, experience, and risk perception. These tests also had mixed results. For instance, self-reported security knowledge was tested in eleven cases, of which five were insignificant, five showed a negative (expected) relationship to susceptibility, and one showed a positive (unexpected) relationship to susceptibility. Concrete tests specifically related to phishing also showed varied results. The variable performance on paper scenarios was related to actual susceptibility in one out of three tests. The tests associated with training included self-reports on training and training provided as a part of the experiment. Significant relationships in the negative direction (expected) were reported in 6of 10 cases and in a positive direction (unexpected) in 1of10 cases. Paired measurements confirmed that training had a good effect overall, but that it was sometimes counterproductive. The relative risk of those receiving training was 0.4 (0.1–1.3) compared to those who did not.

B. The message

This meta-analysis confirms the utility of established deceptive tactics, e.g. that referring to authority or urgency makes a difference. Only two studies addressed such techniques explicitly and the number of samples was smaller than for adaptation. However, the results are clear. On average, cases with an established deceptive tactic obtained a 2.6 (1.8–3.7) higher susceptibility rate. Most of the tests also yielded significant results.

There was little support for the other tested hypotheses related to the message. For instance, in general, it does not appear to be relevant if the purported sender was a female or not, and it did not appear to matter if the email came from the right email server. In addition, it is not clear whether people were less susceptible to emails requiring more risky actions (e.g. providing a password instead of just visiting a web page). It is difficult to establish if this was due to users being ignorant about risks or other (uncontrolled) variables in the studies.

Finally, one message-related aspect not explicitly tested in the studies is whether the content of the scam made a difference, e.g. the storyline in the email. However, this meta-analysis found that this makes a big difference in susceptibility rate. In pairwise comparisons, the better scam yielded a 5.9 (4.3–7.9) times higher susceptibility rate.

C. The situation

Most papers seemed to acknowledge that circumstances related to their experiment were of relevance. For example, it was fairly common to describe how those conducting the experiments ensured that emails reached their destination. There were also more small-scale experiments that showed how browser warnings can reduce susceptibility (relative risk 0.3). However, few studies attempted to actively control or vary the situation in which emails are received. For instance, the time of day, time of week, and time of year that emails are sent were not controlled for, and seldom reported.

A few studies reported results that make it possible to estimate the relevance of situational factors.

The experiment reported in [33] started with a legitimate information email about phishing, in which approximately 40% of the recipients clicked on the link. Those clicking this link in the informational email had susceptibility rates that were higher than the population as a whole (18–19% compared to 14–17%). These numbers suggest that those who deal with more legitimate emails tend to be more susceptible to (i.e. deal with more) phishing emails, which may be due to situational factors (but also possibly due to personality, i.e. recipient factors).

Another study showing how circumstances matter was [55]. Data showed a lot of traffic due to the phishing emails just before closing hours and almost no traffic after closing hours. Thus, timing is of importance.

D. Interactions

This meta-analysis confirmed the widespread notion that *adaptation* matters. Paired measurements with one less and one more adapted variant of a phishing email showed that more adaptation increased susceptibility by an average of 1.5% (1.1–2.0). Significance tests of the variable *personalization* were associated with susceptibility in 5 out of 7 statistical tests (Table II). The adaptations used mostly concerned language, sender and salutation information. One case where susceptibility rates and relative differences were obtained was [38], which tested emails purporting to be a conversational email between the recipient and several of the recipient's friends. A susceptibility rate of 72% was obtained with this email, compared to 16% when purporting to be a stranger. However, there were also a few reports of adaptation decreasing susceptibility. For instance, users were sometimes more susceptible to emails with links to external severs than they were to email with links to internal severs [41], and individualized adaptations (e.g. saluting the recipient by name) were on average no more successful than generic emails (Table I).

One type of adaptation that was largely overlooked in the experiments is the matching between recipients' interests and the content of the phishing email. As mentioned in chapter II.D, one study [39] contained a post-hoc analysis of this matching, finding that emails were seen as more believable when users' work context aligned with that of the email. Other studies also displayed such conjectures. For example, the large difference in susceptibility to the different scams used by [4] may be due to their relevance. As stated by [5, p. 34]: "Messages that targeted issues and concerns relevant to the student sample (e.g., course registration and tuition assistance) were most successful (i.e., in convincing the participants to click the link in the email)." Furthermore, in their post-hoc analysis [4] found that some scams (a gift card and a course registration) were more successful among females. Other post-hoc analyses found that [38] the gender of the sender and receiver matter, with susceptibility being at its lowest when a man receives an email from someone purporting to be a man and increasing for other configurations of sender and receiver.

Finally, [56] tested five hypotheses stating that national culture would interact with another variable (e.g. knowledge) to determine phishing susceptibility. None of these hypotheses were confirmed.

VI. Qualities and flaws in the research

This section summarizes the qualities and flaws in the included field experiments related to phishing.

A. *Descriptions of study context*

The studies on phishing experiments that make up this review were often vague on methodological issues. For instance, it was often difficult to follow the flow of participants, covering each of the stages of enrolment, group assignment, treatment, reception, and analysis. In addition, the papers sometimes referred to other papers as guides for their decisions, but without describing what those guides implied in the study's context. Furthermore, as mentioned in section V of this paper, many of the studies lacked the descriptions needed to classify them according to the classes used in the synthesis of extracted data. Indeed, neglecting descriptions of methodological issues such as experimental design, makes it difficult, if not impossible, to assess the validity of study's result in a certain context.

It would be helpful if researchers explained their reasoning about design choices, to help other researchers evaluate the work, and even to guide other researchers' decisions. For instance, [57] recommended that researchers describe their decisions on how to debrief (or not to debrief) participants. Furthermore, results of statistical tests should be suitably described, e.g. with effect sizes, which would make it easier to make quantitative comparisons between studies ([58]). Different levels of detail could be used in the descriptions of the design, e.g. statements such as "used randomization" and "was double blinded" at the most basic level, and a description of the methods of randomization and blinding at the more detailed level. For instance, the Jadad scale scores clinical trials based on the detail of their methodological quality description [59]. Researchers should even consider providing raw data, since such data could be of great use to other researchers for replication, comprehension, and extension. Admittedly, this poses some problems to the original researchers, e.g. they would potentially be forced to vet each use of the data [60]. Such challenges are similar to those encountered when sharing real (non-synthetic) phishing data with real attackers.

The qualitative criteria used to evaluate the papers included in this review require that the analyzed papers properly describe each part of the research. Indeed, not describing e.g. pilot testing, is treated the same way as saying pilot testing was not conducted. This means that not only more thorough research is rewarded, but to some extent also more thoroughly *described* research. Some authors may take certain things for granted, e.g. that power analysis was performed, and do not see a need to explicitly state this in their paper. On the other hand, more thorough papers, such as theses, have better chances to score highly. Conversely, the criteria used here do not evaluate whether papers were particularly well reasoned (e.g. uses the most topical hypotheses, or drew the most reasonable conclusions from their results). This is worth keeping in mind when reading the rest of the section on qualities and flaws in the research.

B. *Experimental control*

Not all studies formulated clear hypotheses, no studies pre-registered with journals, and few studies were preceded by pilot tests. This is problematic and a potential source for publication bias, i.e. that inconclusive results will never be shared. Power analysis is a useful tool for finding suitable research designs. More specifically, power analysis helps in finding the appropriate (or at least minimum) sample size. Studies that do not use power analysis risk ending up with inadequate sample sizes to find the effect at a particular significance level. A possible reason for not using power analysis may be that the researchers use as large a sample they can achieve. However, even the maximum number available may be too small for the given research, leading to a waste of time and money, and to unnecessary (e.g. ethical) risks. It should be noted that such waste and needless risks may also occur in unnecessarily large sample sizes.

Control groups were far from abundant among the studies, despite control groups being a fundamental part of experimental research. One study described that the studied organization wanted all its employees to receive treatment (i.e. training), which prohibited the use of a control group. Perhaps this is a typical reason. Indeed, there are many stakeholders affecting research, such as researchers, participants, and participant employers ([43], [48]). However, since many authors did not reveal why they did not use a control group, such a conclusion cannot be drawn. It should be mentioned that some studies did not use any treatments, making the need for control groups less of an issue, even though it is rarely possible to avoid any kind of effect on participants (which control groups would be able to compensate for). Beside control groups, the studies did, to a varying degree, mention trying to control for certain things, e.g. whether emails reached their destination or if recipients read them. Further, most studies mentioned checking which participants in the sample who acted on the emails. This lack of control is to some extent to be expected in field experiments. However, it may be the reason for some weak, unexpected, and conflicting results.

C. Ethics and legal issues

While most studies discussed ethical issues, this was typically done rather briefly. IRB approval was only mentioned having been received in 18 out of 48 papers. This is surprisingly low considering the fact that the research typically involves deception of participants and the possibility of harm to them (and to the researchers). Approval by others than IRBs (such as participant employers) was sometimes granted, but it is difficult to see how this can be a substitute for IRB approval, as employers and IRBs typically differ in their ethical stances [52]. Informed consent was virtually non-existent. While realism will suffer if participants know exactly why they are studied, false but plausible hypotheses and areas of interest could be used instead. However, this may sometimes divert the participants' attention, thus also affecting the experiment [62]. An alternative may be to explicitly mention that some sort of deception will occur, akin to not knowing whether a drug or merely a placebo will be distributed [49]. If informed consent is not sought, it would be expected that retroactive consent be pursued after study completion. This would typically take place in combination with the debriefing of participants. However, only 17 papers mentioned debriefing. The aim of debriefing is to remove lies and distress, explain, and allow participants to hold researchers accountable, as well as allowing participants to remove their data (a kind of withdrawal of consent, whether it was ever given) [57]. As with the IRB approval rates, the similarly low debriefing rates are noteworthy, considering the fact that the research typically involves deception and the possibility of harm. Perhaps debriefing was considered too difficult to perform, as the participants were too many and too far from the researchers due to the online nature of the research. This is stated to be a risk by ([49], [52]). It is also possible that debriefing was considered harmful in some cases, due to the difficulty of explaining (especially when performed online and not face-to-face), making participants worried, angry, and potentially initiating lawsuits against researchers [52]. However, [57] instead opined that such arguments are indicative of a study itself being unethical, and that the debriefing only brings that fact to light. Furthermore, if the participants were to learn of the experiment on their own, they could become even angrier and more distrustful than if they were told directly [49].

Some papers mentioned following guidelines on ethical phishing research. Two such guidelines were mentioned and one of these [63] is in fact one of the studies that make up this review. It is rather brief on the topic of ethics, but states that participants ought not to be harmed, that participants must be unaware of their participation during the study, and that striking a balance between accuracy (realism) and the ethical is difficult. As a whole, it is difficult to see why the study should be called a guideline. The other of the two guidelines, [52], contains guiding (but unstructured) statements such as there being a need to have the research approved by an IRB (all four citing papers did), that there is typically a need for deception and an exemption from informed consent, that harm to participants needs to be avoided, and that debriefing can be harmful (as stated above). Further, there are warnings that using company trademarks and intellectual property or violating terms of use may be a problem, and that there are laws against phishing (although they argue that these only apply only to malicious intent). In addition, it is stated that there is uncertainty about what constitutes public data, and how the fact that participants who appear to be of age may in fact be underage and lying about their age. The paper is highly valuable to the discussion of phishing research ethics, but as its authors stated that there is a lack of consensus among IRBs and ethicists, that much more research is needed, and that the paper only gives the outlines for designing and analyzing phishing experiments ethically, it is difficult to see how it can be considered a guideline for such research.

Few papers mention legal issues. While it may be that the approval process included a legal discussion when IRB approval was sought, researchers must take care not to forget legal issues. Such legal concerns may, for instance, include how to deal with the many applicable jurisdictions, e.g. the jurisdiction of the sender, recipient, and those in between [51]. While [51] are not aware of anyone having been found guilty of a criminal or civil offense by a court for conducting academic computer security research, they do recommend that researchers ponder potential risks of legal problems.

VII. Discussion

In this section, the validity and reliability of the review are described. Second, some suggestions for researchers are presented. Third, recommendations for practitioners are given.

A. Validity and reliability of the review

The validity and reliability of the findings reported in this review are contingent on the representativeness of the included studies, the synthesis of their results, and the quality of the primary studies.

The aim was to include all studies meeting the inclusion criteria. Literature searches were conducted using both structured phrases in Scopus and manual searches in Google Scholar. Only one additional relevant study was found from the references in the included papers. Further, the inclusion criteria were straightforward to apply. However, it is reasonable to expect that the search process failed to include everything of relevance. In particular, Scopus has a limited coverage of studies presented on workshops, degree theses, and industry experience reports. In addition, this search procedure will almost certainly have missed any studies not using the word "phishing".

Synthesis of the results was not always straightforward. The grouping of the variables was mostly subjective and done to the authors' best abilities. In addition, it was sometimes difficult to decide which data to extract. Authors sometimes reported several association tests from the same study. In most cases, both reports were extracted because the authors of the primary studies were unclear about which of the tests was the best. In other cases, data reported in the primary studies were considered so prone to bias that it was not extracted for this review. For example, only susceptibility rates of the first and fifth training emails were extracted from the [32], because it was unclear how much training participants was exposed to before each email, and the design was not balanced. The authors of [32] themselves stated: "This creates difficulty for fair evaluation of performance of employees".

The authors of this paper believe subjectivity and bias have been introduced in the process of synthesizing results. This is a

natural consequence of the varied descriptions and design in the reviewed studies, as well as the absence of an established framework for classifying phishing conditions and phishing experiments.

Another issue with the classes used to synthesize the results in this review is that they fail to capture everything of relevance. Within the classes, studies sometimes reported a susceptibility rate that differed substantially and with contradictory associations. Much of this variation could probably be explained if additional experimental conditions were accounted for. However, the varied and incomplete descriptions provided in the papers prohibit such synthesis. For instance, less than half of the included experiments described the amount of time waiting for users to act on a phishing email. Another part of the variation can be expected to stem from natural variability and factors that were unknown to the authors of the primary studies. For instance, it may be that some experiments were biased because participants became aware of the experiment through rumors, and some emails may have been intercepted by spam filters.

Regardless of the reasons for the variation, the results reported in the studies demonstrate that a set of (partly unknown) variables determine peoples' phishing susceptibility. This meta-analysis has tried to determine which these variables are, and nuance the results of individual studies by comparing them to other, similar, studies. The reliability of this analysis is limited by the subjectivity in the classifications, while the validity is limited by the scope of the literature search and the quality of the primary studies.

B. Suggestions for future research

As shown in this review, there have already been plenty of studies on phishing. However, the diverse results of this review also show that the variables that determine individuals' phishing susceptibility are largely unknown. Below are some suggestions for future research on phishing susceptibility.

First, the experimental designs used in the studies are sometimes questionable. For instance, only limited conclusions can be drawn when training is provided only to those who are deceived by a phishing email. Similarly, it is unfortunate when sequences of emails are sent to different participants without any attempt to balance the sequences. It may be that these studies were performed in conjunction with a training program provided by organizations, and that the organizations limit the freedom researchers have to design the procedures. At least one study stated that this was the case [64]. Furthermore, researchers also need to be careful in their relationship to the participants. Ethical issues such as weighting the need for realism with participants' right to informed consent are still debated. Regardless, experimental designs could be more rigorous in many papers, and there are several good examples showing that this is possible.

Second, the research was focused on the recipient, with few tests on how the message and the situation influence susceptibility. For instance, there is a large variation in the effectiveness of different scams, but few studies are designed to identify why some scams work better than others. In addition, email load and how many benign (non-phishing) emails participants act on may be measured. If a higher email load does not lead to participants acting on fewer emails, it is reasonable to assume that the higher load has led to them spending less time evaluating the trustworthiness of the emails, leading to higher susceptibility. However, very few studies described how likely their participants were to act on benign emails, or the extent of the email load. Thus, research on variables related to the message and the situation can be expected to produce new interesting results. For instance, future research may draw inspiration from other kinds of phishing than the ones conducted via email, e.g. text message phishing, giving rise to new types of messages and situations. Combining research on different phishing mediums could provide a more robust stance against future shifts among deceivers.

Third, the aims of the studies could be clearer, going beyond merely determining variables that determine susceptibility rates, and towards reasoning in terms of reducing susceptibility. For instance, concluding that a certain situation or type of recipient personality increases susceptibility should be combined with suggestions on how to deal with this increased susceptibility, e.g. through targeted educational efforts or technical counter-measures. This puts demands on finding all those, and only those, who have the said personality. Likewise, phishing attempts do not only differ in terms of their potential to be realized, but also in the possible consequences if they are realized. A risk-conscious anti-phishing endeavor must combine susceptibility and potential consequences as well as possible counter-measure costs, in order to determine the most cost-efficient steps.

C. Suggestions for practicioners

This review confirms the widespread notion that phishing attacks are a serious threat. The field experiments reported that, on average, 24% of those receiving a phishing email will click on a link and 21% will provide their passwords. These numbers should frighten any system administrator or security manager. Only 2–3% of recipients execute a file provided in a phishing email. However, this should also be considered problematic. In an organization with hundreds of email users, a 2–3% susceptibility rate implies that that the probability of someone running the executable is substantial. In fact, some tests have reported that only 0.6% of server-side exploits work when a network scan suggests that a vulnerability is present [65]. Thus, users can be said to agree with code execution requests more often than machines claimed by scanners to be vulnerable.

This review also finds support for some widespread beliefs concerning interventions that security managers can perform. In particular, training seems to reduce phishing susceptibility, and warnings in technical systems about phishing threats seems to have an effect. As adaptation increases phishing susceptibility, threat intelligence that indicates when the organization or individuals of a certain type are being targeted may be helpful.

The results of this review suggest that peoples' personality has a limited and weak relationship to phishing susceptibility. For instance, it is not clear if those who tend to trust others are more susceptible to email phishing. Thus, procedures that rely on personality traits to determine which users that need more training on phishing are unlikely to be effective.

REFERENCES

[1] Verizon RISK Team et al, "2019 Data Breach Investigations Report," 2019.

[2] Cofense, "The state of phishing defence," 2018.

[3] E. J. Williams, J. Hinds, and A. N. Joinson, "Exploring susceptibility to phishing in the workplace," *Int. J. Hum. Comput. Stud.*, vol. 120, no. July, pp. 1–13, 2018.

[4] S. Goel, K. Williams, and E. Dincelli, "Got phished? Internet security and human vulnerability," *J. Assoc. Inf. Syst.*, vol. 18, no. 1, pp. 22–44, 2017.

[5] A. Ferreira and P. Vieira-Marques, "Phishing Through Time: A Ten Year Story based on Abstracts," in *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, 2018, no. Icissp, pp. 225–232.

[6] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani, "A survey of phishing email filtering techniques," *IEEE Commun. Surv. Tutorials*, vol. 15, no. 4, pp. 2070–2090, 2013.

[7] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Commun. Surv. Tutorials*, vol. 15, no. 4, pp. 2091–2121, 2013.

[8] R. Heartfield and G. Loukas, "A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks," *ACM Comput. Surv.*, vol. 48, no. 3, 2015.

[9] S. Kalra, "A comparative analysis of phishing detection and prevention techniques," *Int. J. Grid Distrib. Comput.*, vol. 9, no. 8, pp. 371–384, 2016.

[10] G. Varshney, M. Misra, and P. K. Atrey, "A survey and classification of web phishing detection schemes," *Secur. Commun. Networks*, vol. 9, no. 18, pp. 6266–6284, 2016.

[11] J. Zhang, C. Wu, D. Li, Z. Jia, X. Ouyang, and Y. Xin, "An empirical analysis of the effectiveness of browser-based antiphishing solutions," *Int. J. Digit. Content Technol. its Appl.*, vol. 6, no. 7, pp. 216–224, 2012.

[12] H. Z. Zeydan, A. Selamat, and M. Salleh, "Current state of anti-phishing approaches and revealing competencies," *J. Theor. Appl. Inf. Technol.*, vol. 70, no. 3, pp. 507–515, 2014.

[13] H. Z. Zeydan, A. Selamat, and M. Salleh, "Survey of anti-phishing tools with detection capabilities," in *Proceedings - 2014 International Symposium on Biometrics and Security Technologies, ISBAST 2014*, 2015, pp. 214–219.

[14] H. Sharma, E. Meenakshi, and S. K. Bhatia, "A comparative analysis and awareness survey of phishing detection tools," in *RTEICT 2017 - 2nd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, Proceedings*, 2018, vol. 2018-Janua, pp. 1437–1442.

[15] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," *Expert Syst. Appl.*, vol. 106, pp. 1–20, 2018.

[16] A. Ferreira and R. Chilro, "What to Phish in a Subject?," in *Financial Cryptography and Data Security*, 2017, pp. 597–609.

[17] E. R. Leukfeldt, "Phishing for suitable targets in the Netherlands: Routine activity theory and phishing victimization," *Cyberpsychology, Behav. Soc. Netw.*, vol. 17, no. 8, pp. 551–555, 2014.

[18] S. Sheng, J. Hong, P. Kumaraguru, L. F. Cranor, and A. Acquisti, "Teaching Johnny not to fall for phish," *ACM Trans. Internet Technol.*, vol. 10, no. 2, pp. 1–31, 2010.

[19] A. Neupane, N. Saxena, K. Kuruvilla, M. Georgescu, and R. Kana, "Neural Signatures of User-Centered Security: An fMRI Study of Phishing, and Malware Warnings," no. February, pp. 1–16, 2014.

[20] B. M. DePaulo, B. E. Malone, J. J. Lindsay, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," *Psychol. Bull.*, vol. 129, no. 1, pp. 74–118, 2003.

[21] R. T. Wright, M. L. Jensen, J. B. Thatcher, M. Dinger, and K. Marett, "Research Note —Influence Techniques in Phishing Attacks: An Examination of Vulnerability and Resistance," *Inf. Syst. Res.*, vol. 25, no. 2, pp. 385–400, Jun. 2014.

[22] J. S. Armstrong, *Persuasive Advertising: Evidence-based Principles*. London: Palgrave Macmillan UK, 2010.

[23] P. E. Johnson, R. G. Berryman, K. Jamal, and S. Grazioli, "Detecting deception: adversarial problem solving in a low base-rate world," *Cogn. Sci.*, vol. 25, no. 3, pp. 355–392, 2001.

[24] K. Marett and R. Wright, "The effectiveness of deceptive tactics in phishing," in *15th Americas Conference on Information Systems 2009, AMCIS 2009*, 2009, vol. 4, pp. 2583–2591.

[25] D. B. Buller and J. K. Burgoon, "Interpersonal deception theory," *Commun. Theory*, vol. 6, no. 3, pp. 203–242, 1996.

[26] T. Halevi, J. Lewis, and N. Memon, "A Pilot Study of Cyber Security and Privacy Related Behavior and Personality Traits," 2014.

[27] G. D. Moody, D. F. Galletta, and B. K. Dunn, "Which phish get caught An exploratory study of individuals′ susceptibility to phishing," *Eur. J. Inf. Syst.*, vol. 26, no. 6, pp. 564–584, 2017.

[28] W. Rocha Flores, H. Holm, G. Svensson, and G. Ericsson, "Using phishing experiments and scenario-based surveys to understand security behaviours in practice," *Inf. Manag. Comput. Secur.*, vol. 22, no. 4, pp. 393–406, 2014.

[29] J. P. Forgas and R. East, "On being happy and gullible: Mood effects on skepticism and the detection of deception," *J. Exp. Soc. Psychol.*, vol. 44, no. 5, pp. 1362–1367, 2008.

[30] C. F. Bond, Jr. and B. M. DePaulo, "Accuracy of Deception Judgments," *Personal. Soc. Psychol. Rev.*, vol. 10, no. 3, pp. 214–234, 2006.

[31] S. Kleitman, M. K. H. Law, and J. Kay, "It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling," *PLoS One*, vol. 13, no. 10, 2018.

[32] H. Siadati, S. Palka, A. Siegel, and D. McCoy, "Measuring the effectiveness of embedded phishing exercises," in *Proceedings of the 10th USENIX Conference on Cyber Security Experimentation and Test*, 2017, pp. 1–8.

[33] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong, "Lessons from a real world evaluation of anti-phishing training," in *eCrime Researchers Summit, eCrime 2008*, 2008.

[34] C. NGUYEN, "LEARNING NOT TO TAKE THE BAIT : AN EXAMINATION OF TRAINING," 2018.

[35] M. L. Jensen, M. Dinger, R. T. Wright, and J. B. Thatcher, "Training to Mitigate Phishing Attacks Using Mindfulness Techniques," *J. Manag. Inf. Syst.*, vol. 34, no. 2, pp. 597–626, 2017.

[36] A. Vishwanath, T. Herath, R. Chen, J. Wang, and H. R. Rao, "Why do

people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model," *Decis. Support Syst.*, vol. 51, no. 3, pp. 576–586, 2011.

[37] H. Holm, W. R. Flores, and G. Ericsson, "Cyber security for a Smart Grid - What about phishing?," in *2013 4th IEEE/PES Innovative Smart Grid Technologies Europe, ISGT Europe 2013*, 2013.

[38] T. Jagatic, N. Johnson, M. Jakobsson, and F. Menczer, "Social Phishing," vol. 2005, pp. 1–10, 2005.

[39] K. Greene, M. Steves, M. Theofanos, and J. Kostick, "User Context: An Explanatory Variable in Phishing Susceptibility," no. February, pp. 1–14, 2018.

[40] J. W. Ragucci and S. A. Robila, "Societal Aspects of Phishing," in *2006 IEEE International Symposium on Technology and Society*, 2006, pp. 1–5.

[41] B. M. Bowen, R. Devarajan, and S. Stolfo, "Measuring the human factor of cyber security," in *2011 IEEE International Conference on Technologies for Homeland Security, HST 2011*, 2011, pp. 230–235.

[42] S. A. Olivo, L. G. Macedo, I. C. Gadotti, J. Fuentes, T. Stanton, and D. J. Magee, "Scales to Assess the Quality of Randomized Controlled Trials: A Systematic Review," *Phys. Ther.*, vol. 88, no. 2, pp. 156–175, 2008.

[43] A. Bryman, *Social Research Methods*, Fourth. Oxford University Press Inc, 2012.

[44] D. Muise and J. Pan, "Online field experiments," *Asian J. Commun.*, vol. 29, no. 3, pp. 217–234, 2019.

[45] J. A. Konstan and Y. Chen, "Online Field Experiments : Lessons from CommunityLab," in *Proceedings of Third International Conference on e-Social Science*.

[46] W. W. C. Q. Review, H. Start, I. Study, and F. Report, "What Works Clearinghouse," *Education*, no. December, 2010.

[47] J. C. Valentine and H. Cooper, "A Systematic and Transparent Approach for Assessing the Methodological Quality of Intervention Effectiveness Research: The Study Design and Implementation Assessment Device (Study DIAD)," *Psychol. Methods*, vol. 13, no. 2, pp. 130–149, 2008.

[48] D. R. Ilgen and B. S. Bell, "Conducting industrial and organizational psychological research: institutional review of research in work organizations," *Ethics Behav.*, vol. 11, no. 4, pp. 395–412, 2001.

[49] D. B. Resnik and P. R. Finn, "Ethics and Phishing Experiments," *Sci. Eng. Ethics*, vol. 24, no. 4, pp. 1241–1252, 2018.

[50] K. Jüristo, "How to Conduct Email Phishing Experiments," 2018.

[51] C. Soghoian, "Legal risks for phishing researchers," *eCrime Res. Summit, eCrime 2008*, 2008.

[52] P. Finn and M. Jakobsson, "Designing ethical phishing experiments," *IEEE Technol. Soc. Mag.*, vol. 26, no. 1, pp. 46–58, 2007.

[53] A. Stephanou, "The Impact of Information Security Awareness Training on Information Security Behaviour," University of the Witwatersrand, 2008.

[54] W. Yang, A. Xiong, J. Chen, R. W. Proctor, and N. Li, "Use of phishing training to improve security warning compliance: Evidence from a field experiment," in *ACM International Conference Proceeding Series*, 2017, vol. Part F1271, pp. 52–61.

[55] T. Bakhshi, M. Papadaki, and S. M. Furnell, "A practical assessment of social engineering vulnerabilities," in *Proceedings of the 2nd International Symposium on Human Aspects of Information Security and Assurance, HAISA 2008*, 2008, pp. 12–23.

[56] W. R. Flores, H. Holm, M. Nohlberg, and M. Ekstedt, "Investigating personal determinants of phishing and the effect of national culture," *Inf. Comput. Secur.*, vol. 23, no. 2, pp. 178–199, 2015.

[57] R. Sommers and F. G. Miller, "Forgoing Debriefing in Deceptive Research: Is It Ever Ethical?," *Ethics Behav.*, vol. 23, no. 2, pp. 98–116, 2013.

[58] D. Moher, K. F. Schulz, and D. G. Altman, "The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials," *Lancet*, vol. 357, no. 9263, pp. 1191–1194, Apr. 2001.

[59] A. R. Jadad *et al.*, "Assessing the quality of reports of randomized clinical trials: Is blinding necessary?," *Control. Clin. Trials*, vol. 17, no. 1, pp. 1–12, 1996.

[60] D. R. Thomas, S. Pastrana, A. Hutchings, R. Clayton, and A. R. Beresford, "Ethical issues in research using datasets of illicit origin," in *Proceedings of the 2017 Internet Measurement Conference on - IMC '17*, 2017, pp. 445–462.

[61] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor, "The preregistration revolution," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 11, pp. 2600–2606, 2018.

[62] A. I. Piper, "Conducting social science laboratory experiments on the world wide web," *Libr. Inf. Sci. Res.*, vol. 20, no. 1, pp. 5–21, 1998.

[63] M. Jakobsson and J. Ratkiewicz, "Designing ethical phishing experiments: A study of (ROT13) rOnl query features," *Proc. 15th Int. Conf. World Wide Web*, pp. 513–522, 2006.

[64] R. M. Long, "Using Phishing to Test Social Engineering Awareness of Financial Employees," 2013.

[65] H. Holm and T. Sommestad, "So long, and thanks for only using readily available scripts," Emerald Group Publishing Ltd., Mar. 2017.