The Role of Overconfidence:
Getting Individuals to Recognise Phishing Emails

23 February, 2023

ING Belgium
Workgroup EBB2a

Zeynep Bagatir 658367
Gijs Harmeling 619524
Jackson Kent 637817
Ella Needler 592572

Word count: 8744

Executive summary

Introduction and Solution:

Cyber attacks are very common in the digital era. Even though a lot of people fall for scams such as phishing, individuals are not yet protecting themselves sufficiently. We believe that a reason for the lack of prevention against phishing is that people are overconfident in their capacities to recognize these emails, so they think it will not be them who fall victim. Our solution to this issue is to confront individuals by having them experience failure.

Methods:

The way we tested our solution is through a survey. The survey contained 15 emails, both genuine and phishing. The control group simply went through the emails, but the treatment group got an intervention after 5 emails, which showed them how poorly they performed up until then. If there would be a difference in scoring in the last 10 emails between the two groups this would indicate that confrontation with their abilities could increase an individual's ability to recognize phishing emails.

Findings:

Our results indicate that the majority of individuals were indeed overconfident. Furthermore, the intervention reduced confidence levels of the treatment group more than the control group, but this difference was insignificant. Also, when looking at a binary indicator for the dependent variable, we see that the score increased a lot more for the treatment group, yet this was not significant either.

Implementation and evaluation:

Based on previous literature and the indications of our results, we believe that ING should send-out "mock attacks" to its customers, to confront the individuals after they click on this fake email. To avoid upset customers they could first select participants through an opt-in program. Evaluation of this solution could be done through continuous sending of fake emails as well as having participants do a test to check their ability to recognize phishing.

Introduction

In the Netherlands around 2.5 million individuals, aged 15 and up, fall for cyber crime annually (CBS, 2022). In Belgium the problem is even greater with over 40% of the population being a victim of phishing in 2022 (Brussels Times, 2022). Clearly, online crime and phishing form a serious problem, but what exactly does it entail? Phishing is an illegal activity where fraudsters obtain personal and financial information through electronic communication tools such as text messages or emails. It is a well-known cybersecurity attack that has become more common in recent years. It poses risks to businesses, government agencies, and all users due to sensitive data breaches and subsequent financial losses. Financial institutions are one of the most affected sectors by this increase. With the widespread use of online banking in recent years, cybersecurity concerns have also increased. In particular, banks are having problems to prevent all phishing incidents. Hong (2012) suggested that the institutional efforts alone are not enough to eliminate phishing attacks, the individuals who are the target of these attacks need to also be wary. The solution to the cybersecurity problem lies with the individuals. Companies should focus on people instead of just investing in technology (Hewitt & White, 2022). A clear indication of this is the case presented by ING Belgium which is a financial institution in Belgium with 3 million customers. Despite all the efforts of the bank to raise awareness, their customers cannot avoid being victimised by online fraud. Although everyone is aware of such fraudsters and their techniques, users take no precaution against the fraudster and they keep falling into their trap. Therefore, it can be concluded that the users' behaviour does not rely on a rational decision making process (Metzger & Suh, 2017). Our study will focus primarily on the user side and try to explain why users do not care enough about their own cyber security. To study the user side, this paper aims to conduct a literature review and user study.

ING stated that one of the most commonly used phrases by the victims is "I thought I knew". This sentence might be the key to understanding the behaviour of users, since it shows that overconfidence bias might be the driving force behind their misjudgements. Overconfidence bias is a cognitive bias that refers to the tendency of people to have a misjudged perception about the abilities they have. In this case users assess their ability to recognize phishing higher than is really the case. Therefore, if the goal is to minimise the number of fraud victims, a way should be found to reduce the overconfidence bias. One way to achieve this goal might be to show users that they are actually not as good as they thought, and let them

experience failure. Hence, the hypothesis that will be tested is if people who are confronted with their failure to recognize phishing emails are more likely to recognize phishing emails in the future. To be specific, the question that this paper will try to answer is to what extent experiencing failure reduces overconfidence bias and increases the ability to recognize phishing emails. Although none of our results were significant, there are some indications that overconfidence could be reduced by confrontation. Our research showed that indeed more than 60% of the participants were overconfident and that being in the treatment group did reduce this confidence more than for people in the control group. Based on this and previous literature, we suggest that ING uses "mock attacks" to test its customers who choose to opt-in to the programme. If customers fail and click on the link in these fake phishing emails, they will then be confronted with their lack of knowledge. By comparing the results over time, it should then become clear that individuals reduce their confidence and increase their accuracy after such an experience.

Literature Review

As explained in our introduction, phishing can be the cause of data breaches, compromised credentials and significant financial loss. As a result it is relevant to understand why people fall for these attacks and quite some research has already been devoted to this. Apart from all the technological innovations to increase cyber security, such as security indicators and appearance of website URLs, researchers have used a behavioural approach to understand the mental processes of why people fall for phishing emails (Das, Camp & Nippert-Eng, 2022). This is particularly relevant as the initial judgement of an email will largely determine the reaction of the receiver (Wang, Li & Rao, 2016). That humans are the most important part in the process of phishing was already mentioned by Hong, who stated that "*It doesn't matter how many firewalls, encryption software, certificates, or two-factor authentication mechanisms an organisation has if the person behind the keyboard falls for a phish*" (2012). This is why we also emphasize that organisations should not spend their time and resources on merely the technology, but on motivating people and influencing their behaviour as well (Hewitt & White, 2022).

To understand how to change people's behaviour we have to start with the fact that people are bounded by rationality, i.e. people are not rational decision makers (Simon, 1986). More concretely, in our case there seems to be an inconsistency between the knowledge of people about the risks of phishing and the precautionary actions they take (ING, 2023; Metzger and

Suh, 2017). People put too much trust in other parties when being active online and, especially when people are under stress, they fail to make rational decisions (Kumaraguru, 2007). These characteristics are very common in phishing emails, where the scammers pretend to be a trusted organisation and pressure the victim into making quick and irrational decisions.

The poor choices made by people in these circumstances can be explained by the division of System 1 and System 2 thinking, first proposed by Daniel Kahneman (2011). Before Kahneman, Michael Posner et.al, already identified two types of thinking, identifying them as "automated" vs "controlled" (Posner, Snyder & Solso, 2004). System 1 is thought to be a quick, intuitive process, which happens almost automatically with little effort. It is primarily driven by instinct and experience, while System 2 thinking requires more effort and is slower. As a result, System 2 thinking is conscious and more logical. For our case, we expect that more cognitive effort, such as increased use of System 2 thinking, would make receivers of phishing emails more likely to recognize these as such. In fact, research has shown that rapid judgments can cause decision errors, as individuals are not taking into account a wide range of relevant cues. Increasing the effort when processing emails, can make individuals more capable in noticing the abnormalities of phishing emails and as such be less susceptible to them (Wang, Li & Rao, 2016). Parsons et. al (2013) found that participants who obtained a higher score on the test of cognitive impulsivity, so the respondents who could control their impulsivity better, were significantly better at identifying phishing emails. Thus, the respondents who probably deliberated more over a phishing email appeared to respond better (Kumaraguru et al., 2007).

Indeed the constantly improving and increasingly cunning technologies and methods used for phishing, which become increasingly deceptive, require recipients to be more cautious for different aspects of the phishing mail. To find the phishing mails will require extensive effort. Still, the opposite happens. People use their email a lot and experience a lot of (time) pressure in dealing with all the information they receive, which causes the System 1 thinking to take the lead. As a result people do not pay enough attention when examining emails, even ignoring indicators as the actual addresses in the mails, and will thus misjudge (Dhamija et al., 2006; Pattinson et al., 2012). One reason why people have little decision effort when judging these mails is overconfidence bias in decision making (Kassin et al., 1991).

Confidence has proven to play a critical role when predicting human behaviour (Fazio & Zanna, 1978). Moreover, confident individuals will also act consistently with their beliefs about their judgement. Previous research on confidence and phishing showed that around 92% of their participants misclassified phishing emails even though 89% indicated they were confident of their ability to identify phishing emails (Hong et al., 2013). Reasons for this overconfidence could be the perceived familiarity individuals have with a certain brand, their bank or for example the government (Alba & Hutchinson, 2000), or because they simply think it will not happen to them, which we will elaborate on later. If indeed an individual is overconfident, when their confidence exceeds their performance, this could lead to risk prone behaviour (Moore & Healy, 2008). For phishing, this could mean that people will not properly check an email they judge to be a genuine email, because they are so confident they would recognize phishing emails immediately. In reality, this person is more likely to respond to the email, release personal information and lose credentials or money. If an individual has lower confidence, this person is more likely to avoid these consequences by putting more effort in verifying the authenticity of the email and its sender. In this way overconfidence may indeed have the results that people take actions they should not have, which leads to poor and even disastrous outcomes (Tang et. al, 2014).

The relevance of overconfidence in phishing attacks is further elaborated by Wang et al. (2016: 760), "*This illustration of judgmental confidence and the issue of overconfidence in phishing detection implies that a better correspondence between confidence and accuracy would help prevent one from falling to phishing*". The results of Wang et. al. (2016) show that in a broad demographic sample, people are overconfident in their capabilities of phishing detection. And they argue that both research and practice need to pay more attention to this issue. Moreover, they show that overconfidence was decreased by cognitive effort, which is once more a confirmation that getting people into their System 2 rather than System 1 thinking could be part of the solution to fight phishing.

Related to the problem of overconfidence is the optimism bias. Literature suggests that optimism can also lead to relying on biases and heuristics which can cause overconfidence (Hayward, Shepherd & Griffin, 2006), in turn this can result in overconfidence in judgement under uncertainty (Libby & Rennekamp, 2011). The optimism bias refers to the fact that humans tend to overestimate the likelihood of positive and underestimate the likelihood of negative events. This is also a primary reason why people perform actions that are currently

rewarding, but could be costly in the future, such as smoking. Weinstein et al. (2005) proved that when assessing their risk of lung cancer smokers demonstrate an optimism bias. This bias is said to be one of the most prevalent and robust biases in psychology and behavioural economics (Sharot, 2011). Findings have shown that the optimism bias is also a problem for cyber security (Hewitt & White, 2022). Even when people are aware of the risks posed by phishing, they can hold an unfavourable attitude toward taking preventive measures because of their optimism bias. We believe that the optimism bias is fairly similar to the overconfidence bias in the sense that they both lead internet users to believe that it will not be them who will be the victim of phishing.

We believe overconfidence bias and the optimism bias are key mechanisms in understanding an individual's behaviour towards phishing scams. Therefore, simply providing information or educating people through awareness on the dangers of phishing is not a sufficient solution as people continue to keep their overconfident beliefs. Research shows security awareness is a poor solution for improving an individual's security behaviour (Goel et al., 2020; Yoon et al., 2012). In a study to understand educational methods for increasing students' security compliance, Yoon et. al. (2012) conclude hand-ons learning as the most effective solution. Therefore, we will take a different approach than educational awareness to fighting these biases: experiences. For our research we will try to find a way of confronting individuals with their poor performance on detecting phishing mails and we will assess whether this confrontation influences the capabilities of these individuals. The research question we will try to answer in this regard is:

> *To what extent does experiencing failure reduce overconfidence bias and increase ability to recognize phishing emails?*

We hypothesise that after being confronted with how little knowledge an individual has about phishing and recognizing phishing emails, people will be less confident in their capabilities and more likely to understand that they too can be a victim of phishing. As a result, we expect that experiential learning will lead them to engage their System 2 thinking and thus be more conscious when having to judge future emails on their genuineness. This leads to our hypothesis:

> *People who are confronted with their failure to recognize phishing emails are more likely to recognize phishing emails in the future*.

In the next section we will further explain our solution and elaborate on the theoretical reasoning behind this solution.

Solution

Whenever people have to make decisions under uncertainty, they often use confidence as an indicator of their accuracy and as such it plays a very significant role in the decision making process of individuals (Camerer & Lovallo, 1999; Chuang & Lee, 2006; Hirshleifer & Luo, 2001). In uncertain situations, only after an event has occurred, people will be able to see the actual result. In the case of phishing: the loss of information or money. This means that when people are overconfident, confidence in fact is a poor advisor, and could lead them to make the wrong decision. Indeed, as discussed before, research has shown that overconfidence is a problem in phishing email detection. Therefore, mechanisms should be created to mitigate overconfidence and enhance judgement. This is exactly what we propose to do with our solution.

There are various solutions that have been tried to stop people from falling for phishing schemes, from technological solutions to informing and educating internet users. Indeed these could be helpful, because to deal with overconfidence awareness of the problem is always needed; only after people understand the risks of phishing attacks, will they take countermeasures (Lei, Hu, & Hsu, 2022). Still, at the moment people do not feel the risk, as it was shown that people who knew more about information systems or technology, were less likely to recognize phishing emails. Thus it is probably the complacency, or overconfidence, of individuals that needs to be appointed directly (Parsons et. al, 2013). The goal of our solution is to make people stop and think about how to respond to an email, instead of them trusting their own capabilities. This is confirmed by training literature, which tells us that changing behaviours is better than merely teaching people rules (Parsons et. al, 2010).

Our solution will try to fight the belief of individuals that they will not be the ones falling for a phishing scheme, because of their optimism bias or overconfidence, and confront them with the privacy risks posed by phishing. We believe this to be the most effective countermeasure to convince people that they are equally susceptible to the risks as anyone else. Simply providing information or educating people with awareness will not be sufficient, as people will continue to keep their overconfident beliefs that they know better and that they will not be part of the victimised group. In other words, users are unmotivated to read about cyber

security and do not learn how to protect themselves. Therefore, our solution will be to have these overconfident individuals experience failure themselves and confront them with the little knowledge they have on recognizing digital threats. In the following we will further explain this.

As mentioned before, simply sending information and instruction materials is not a great way of motivating people to spend time on understanding the material. Mainly because people do not understand why they receive the emails, they simply delete them or leave them unread. On the contrary, if people fall for a phishing email, they understand the need to pay more attention (Kumaraguru et. al., 2007). In their research Kumaraguru et. al. found that people learn much better after falling for an attack than when exactly the same training materials are sent via email. This is in line with the competence hypothesis of Heath and Tversky (1991). This hypothesis states that individuals feel more confident in a situation where they understand what is happening and they feel knowledgeable or competent than in a situation where they have less knowledge or feel ignorant. They argue that general knowledge, experience, and familiarity can all enhance an individual's feeling of competence and indeed phishers will play into these characteristics that create a sense of competence. This is why being confronted with their own ignorance could have a strong effect on these individuals, reduce overconfidence and create a sense of urgency to better recognize phishing emails.

Therefore, as solution, we propose the following:

> *Users should be sent simulated phishing attacks to be confronted with their lack of ability to recognize these attacks.*

If they indeed fall for one of these mock attacks, they will not lose their private data or money, but they will lose part of their confidence in recognizing phishing emails. In the end this should lead them to pay more attention and more consciously analyse future emails, as failing creates a much stronger motivation for these users. One of the reasons this works is because it applies the learning-by-doing and immediate feedback principles.

The learning-by-doing principle comes from the adaptive control of thought-Rational (ACT-R) theory of cognition and learning and it refers to the fact that knowledge is easier acquired and strengthened by doing, through practice (Anderson, 1993). With our solution, users experience failure and should learn through this experience. The feedback principle is

the other part of the solution. People need to receive their results immediately after performing the task. Arkes et al. (1987) gave subjects practice problems and then provided intermediate feedback. Through this they managed to reduce overconfidence and the results of the people who received this feedback improved significantly for the remainder of the test.

Russo & Schoemaker (1992) stated already that predictions will almost certainly be overconfident, but good feedback will quickly reduce it. Feedback makes people more aware of their optimism and overconfidence bias (Moores & Chang, 2009; Sharp et al., 1988). For this reason, we confront people with their score compared to the confidence level they report at the beginning of the survey. Giving immediate feedback provides guidance towards the right behaviour. (Schmidt & Bjork, 1992). In our research we will give participants feedback after they answered five questions, but this will be slightly different for the actual solution to be implemented by ING. We will elaborate on both later.

The main criterion for our solution to work is that individuals must change their view on their capabilities to detect phishing emails. The first step in the process of reducing the overconfidence is to confront people with how little they know. This goes hand in hand with our first constraint as well. It might be the case that people simply do not fall for the mock attacks or do not respond to them. Obviously, they cannot be confronted with failure then. A second criterion is that people need to receive feedback. The feedback on their performance is what should trigger something with the individual. Being told what they do not know is a critical part of the confrontation process. Finally, people need to get a new chance to recognize a phishing email. This is when we will know whether something actually changed within the mindset and thus the performance of this individual. Some constraints to this solution are the following. It could be that overconfidence is not the only mechanism at play, but that there are various biases at work at the same time. If that is the case, simply reducing overconfidence will not yet solve the problem. Second, the question remains whether simply confronting someone with their lack of capabilities once is enough to reach a long-term effect; habit formation is usually done over a longer period of time. Especially if this confronting experience is not severe enough, which could be an additional constraint, the effect might not be significant and enduring. These are limitations we foresee for our research and solution. We will expand on these in our discussion after presenting our results.

To conclude, let us summarise our solution to fight the overconfidence and optimism bias, which often lead internet users to underestimate the risks of cybercrimes. To make people really feel that they are susceptible to these risks, we need to confront them with how little they actually know about recognizing phishing emails. By using feedback, we will try to reduce the overconfidence and optimism bias and have the subjects learn from their experience. Providing this feedback will likely cause individuals to adjust their confidence downwards and this in turn will, because subjects are more conscious about the choices they make, increase the accuracy of phishing email recognition.

Research Methods

The objective of this study is to investigate the impact of prior experience with falling for a phishing scam on an individual's ability to avoid future attacks. The study uses a survey-experiment design that simulates the effect of falling for a phishing scam to drive down the overconfidence a given subject may have in their own ability to detect phishing emails. To achieve this, the experiment incorporates primary design features from two past studies: one study that measures a subjects' ability to detect phishing scams, and another that reduces overconfidence in its subjects (Pattinson et al. 2012 and Arkes et al. 1987, respectively). Our experimental design draws from these two studies in tandem to examine how reducing overconfidence in our subjects can alter their ability to detect phishing emails. Implementing these empirically backed methods fortifies our study with confidence that our methods truly aim to answer the research question at hand.

A single questionnaire was devised for this study using Qualtrics survey software, where participants were randomly allocated to either a treatment or control group. The allocation was set such that an even number of participants were placed in each group. The main difference between the two groups, our intervention, was that the treatment group was confronted with their poor results after answering the first 5 questions. The questionnaire is composed of four main sections for each group: a preliminary message, a phishing detection task, demographic information, and a debriefing at the end.

The preliminary message serves to inform participants of their task and provides background information, while masking the explicit focus on phishing scams. One main critique of phishing IQ studies is lack of real-world validity. In previous studies where subjects are

informed that they are undergoing a phishing study, they often become biassed in the sense of being more careful to look out for "phishing" indicators (Anandpara et al., 2007). Since this level of suspicion may not be present when the individuals routinely check their inbox in a real-world scenario (Furnell, 2007), some researchers have questioned whether phishing IQ tests can be reliably used to test how susceptible people are to phishing emails. Our study overcomes this drawback by adjusting the context of the questionnaire such that the subject is tasked with merely "organising" the mailbox of a fictitious character "Jack Johnson" as they see appropriate, instead of explicitly partaking in a "phishing detection task." This framing aims to mask the true nature of the study, in hopes that the subject is not primed towards thinking that they need to be extra careful for phishing emails. The subject is then asked if they are above the age of 18 and if they consent for their answers to be used, and are then asked how confident they are in being able to fulfil the task on a scale of 0-100. See Appendix 1 for the initial prompt asking the participant to organise the inbox and their confidence level.

Following these preliminary questions is the phishing detection task, the answers of which will be used as the dependent variable for analysing and answering our research question. Subjects in both treatment and control groups are presented with a set of 15 identical emails, some of which are genuine emails, others phishing emails. Subjects go through each individual email and indicate which of the following actions they would take:

1.) Leave in inbox and address immediately
2.) Leave in inbox
3.) Delete
4.) Delete and block the sender

The answers are scored accordingly for a genuine email, and are inverted for a phishing email:

1.) +2
2.) +1.5
3.) +1
4.) +0.5

This is so that selecting option 1 while looking at a genuine email yields the most points for genuine emails, and the lowest amount for phishing. This scoring is flipped when the participant is viewing a phishing email, where they receive +2 points for clicking option 4

"Delete and block the sender". These answers and scoring systems were originally used by Pattinson et al. (2012) to test a subject's ability to detect phishing emails, and are used for the same purpose in the present study. Answers ranging 1- 4 provide for more heterogeneity in the answers, as opposed to a binary indicator that would just mark "Leave in inbox" or "Delete." The extra options help to explain part of someone's rationale behind the positive or negative action with the email. "Deleting" but not blocking the sender may indicate that the user is unsure about what the best action actually is, and could be left vulnerable to future scam attacks. Nonetheless, in addition to using this scale for our analysis, we also performed our analysis with a binary measure. We did this to make sure that this would not lead to different results, as it could be the case that people were overly cautious and would therefore only choose the middle options. For this binary measure we combined options 1 and 2 and options 3 and 4.

The genuine emails used were collected from the researchers' own inbox, with sensitive information censored. The phishing emails used were provided by ING, and were translated to English for use in the survey. Since the phishing emails needed to be translated in a Word document, all of the emails used were formatted in Word to look similar to a real inbox format, and were screenshotted and presented to the subjects as "hypothetical." Since the emails were done as screenshots and participants could not engage or look at links, we added the genuine URLs to the genuine emails and created fake URLs for the phishing emails where users are typically asked to click. The reason is the core of a URL can be one of the primary ways in detecting a phishing email and we did not want to inhibit the users' ability to distinguish emails this way solely because the survey was using screenshot images of the emails rather than actual emails. There were 15 emails in total. Amongst the set of first 5 emails, 2 were genuine and 3 were phishing. Amongst the set of the last 10 emails, 3 were genuine and 7 were phishing. Examples of a genuine email and phishing email used can be found in Appendix 2.1 and 2.2, respectively.

Now let us elaborate on the treatment. The questionnaire for the treatment group includes feedback on the accuracy of the first five questions immediately following the fifth email, while the questionnaire for the control group does not include such feedback. See Appendix 3 for examples of the intervention participants could receive based on their confidence and accuracy. This method comes from the Arkes et al. (1987) study that was used to test overconfidence reduction techniques in a similar capacity by observing how well subjects

performed on a set of questions before and after an "ego-check." The dependent variable used for our analysis is the score of the second set of questions (the last 10), with the independent variable of interest being a binary indicator of whether the subject was in the treatment or control group. The hypothesis is that the treatment group, having received feedback on their performance, will perform better on the latter ten questions after having been shown their initial score, when compared to the control group. To finish off the email management task, after both groups went through the fifteen emails, they were again asked how confident they were that they managed the inbox appropriately.

Following the email management task is the demographic information section, which primarily allows for sample subsetting for analysis purposes. If this section had come before the phishing task, it is possible that the subject would lose some interest in the survey while filling it out and lose the required focus to complete the main task. Hence, demographic information comes after the phishing task. The remaining questions asked for age, gender, and education level. Finally, the subjects were debriefed on what the experiment was actually about, and how their answers would be used to test if overconfidence in fact played a role in how well they appropriated the emails (see Appendix 4).

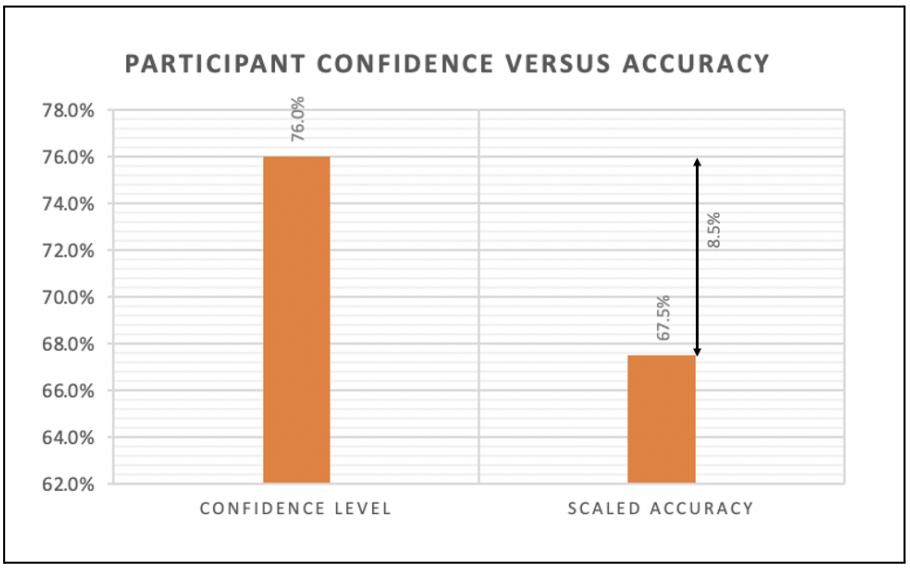
Before starting our data collection we performed a power calculation using STATA. We did the power calculation based on the amount of points a participant could score in the last 10 questions, as that score formed our dependent variable. Input assumptions were based on previous literature, explanations found in Appendix 5. Combining all these numbers we would have to get a sample size of 506 with 253 observations in both the control and treatment group. In the following section, we will go over the results of our survey.

Results

From our survey, we collected a total of 102 observations of which 96 were working observations. 6 observations were removed because they were not completely filled out or were completed in less than 30 seconds. There are 48 working observations in the control group and 48 in the treatment group. Our survey population is distributed fairly evenly amongst gender and age with 43 females, 49 males, and 4 who preferred not to say, as well as, individuals' ranging from age 20 to 69 years old with the majority aged 23-26. Throughout the analysis, the median was used over the average to aggregate results due to the

presence of a few outliers. In addition, both the scaled accuracy, the points system ranging from 0.5 to 2, and the binary accuracy, a simple did they chose to delete or leave the email, were calculated. For the majority of this section scaled accuracy is presented, because in most cases there were only minor differences in results based on the different scoring and given the argument given by previous literature for the scaled accuracy which is explained in the Methods section.

Figure 1:



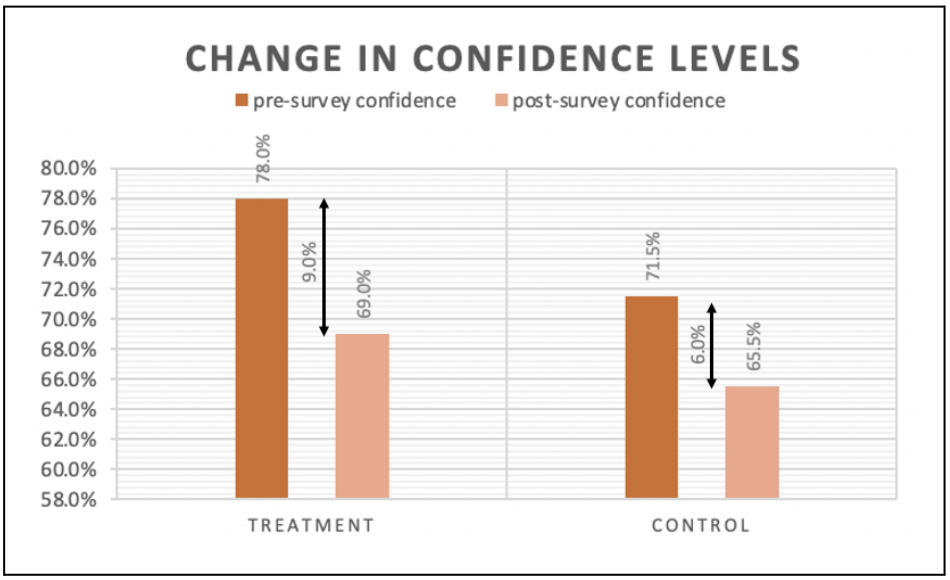
Participant's confidence level at the beginning of the survey compared to the participants accuracy on the first 5 emails.

Given our hypothesis and solution, the main mechanism we want to test for is overconfidence. Thus we need to see if overconfidence is present in the sample. Figure 1 compares a participant's confidence level at the beginning of the survey to their accuracy on the first 5 emails. Looking at these two scores allows us to determine if a participant is overconfident, i.e. if their confidence level outweighs their ability to accurately manage the email inbox and thus detect phishing emails. We see the median confidence level ranked by the participants at the beginning of the survey is 76.0% while the participants median accuracy score is 67.5%. With a difference of 8.5%, our participants overstated their ability to manage the inbox. Moreover, if we look at the participants individually, a total of 58 (or 60.4%) overstated their confidence by ranking a higher confidence score than their accuracy on the first 5 questions. This is fairly evenly distributed across both the treatment and control

groups with 28 and 30 people in each overstating their confidence, respectively. So, we can assert that overconfidence is present amongst our sample.

Next, it is worthwhile to know how participant's confidence changed as they took the survey. Looking at how participant's confidence changed, we see a decline. Figure 2 shows the change in the median confidence levels of both the treatment and control groups from the beginning to the end of the survey. In the treatment group, the median confidence level at the beginning of the survey was 78.0% and at the end 69.0% meaning a 9 point decline in confidence levels from the beginning to the end. In the control group, the median confidence level at the beginning of the survey was 71.5% and at the end 65.5% meaning a 6 point decline in confidence. In both the treatment and control groups, confidence declined.

Figure 2:



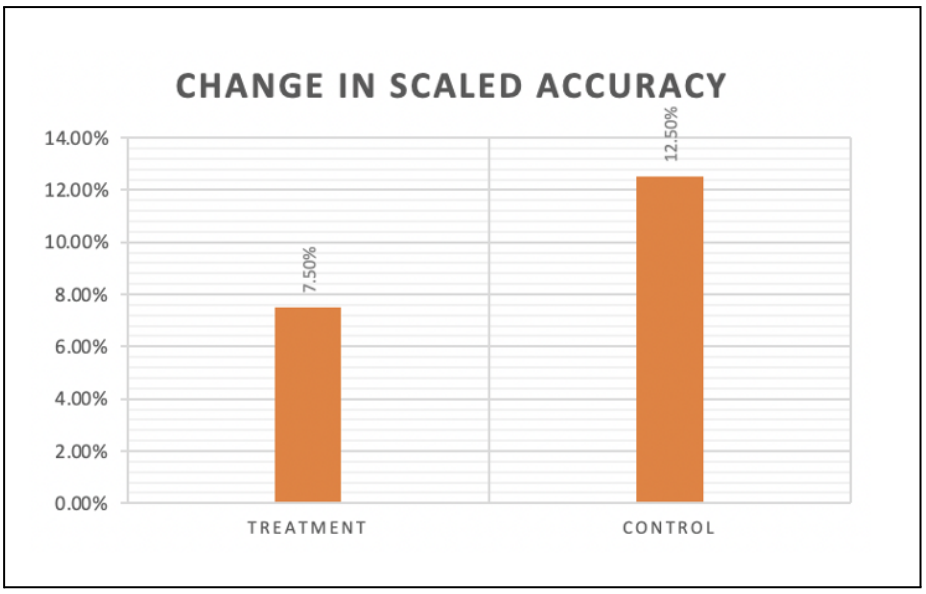
Confidence level indicated by the participant in the beginning of the survey and the end of the survey for the treatment and control groups. The black arrow indicates the change in confidence levels from beginning to end of the survey.

Moreover, there is a 3 point difference in the change in confidence levels between the treatment and control groups, suggesting our intervention may have been effective. To determine if this is statistically significant, a Mann-Whitney U test is conducted to see if the change in confidence levels between the treatment and control groups is statistically significant. The test returns a p-value of 0.814 which is not less than 0.1 so the decline in

confidence level between the treatment and control is not statistically significant. It should be noted that one possible reason for the insignificance is the unbalanced confidence levels for the beginning of the survey. Our survey randomly assigned individuals into treatment and control groups, however, this does not guarantee an even distribution of confidence levels. The pre-survey confidence intervals for the treatment is 78.0% while it is only 71.5% for the control group. While we do look at the change in confidence levels and not the final confidence level, the unbalanced data could still impact the significance.

To continue, if we look at accuracy specifically we see an overall increase in participant's ability to properly manage the inbox for both phishing and genuine emails. The treatment group had a median accuracy of 70.0% in the first 5 emails and a median accuracy of 77.5% in the last 10 emails, meaning there was a 7.5 percentage point increase in accuracy. In the control group, participant's had a median accuracy score of 65.0% in the first 5 emails and 77.5% in the last 10 emails, meaning a 12.5 point increase in accuracy. Figure 3, below, shows the change in scaled accuracy for the treatment and control between the first 5 emails and the last 10 emails. A Mann Whitney U test comparing the accuracy change in the treatment group to the control group returns a p-value of 0.641 and so is statistically insignificant.

Figure 3:

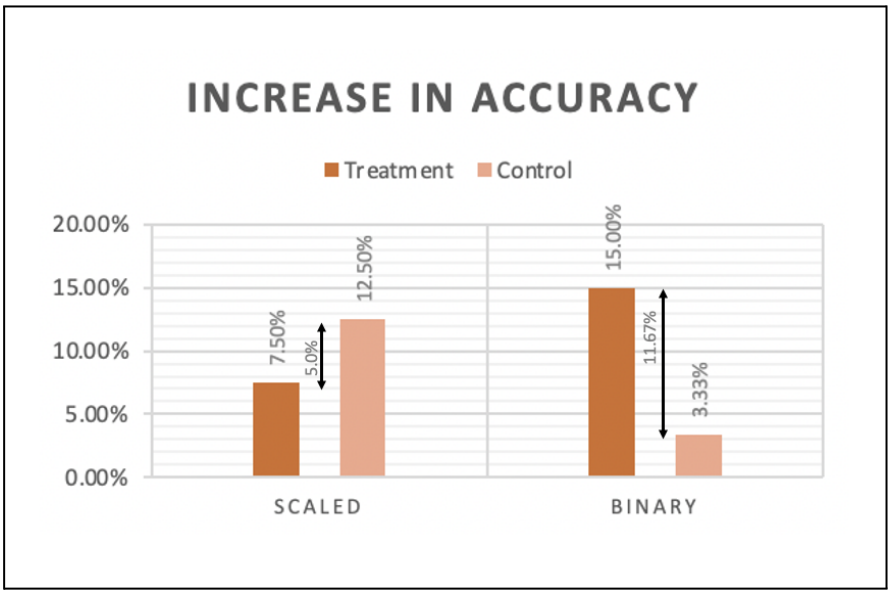


Change in participants accuracy from the first 5 emails to the last 10 emails between the treatment and control groups. The accuracy is scored based on the scaled system explained by Pattinson et al. (2012).

For the sake of robustness, we also looked at overall accuracy with the binary scoring. With binary scoring, we also see an increase in accuracy between the first 5 emails and last 10 emails. In the treatment group, accuracy increased 15 points from 60.0% to 75.0% and in the control group accuracy increased 3.33% from 66.67% to 70%. In the binary scoring, we see the treatment group faces a much larger increase in accuracy than the control by almost 12 percentage points. This can translate to detecting 2 additional phishing emails. To see if this difference is statistically significant, a Mann Whitney U is conducted and shows the difference in the binary ranking is also statistically insignificant with a p-value of 0.737, similar to the scaled accuracy. Although the test is not significant, we still think that the 12 point difference gives an indication that the treatment could have some effect.

Figure 4, below, shows the change in accuracy for both the scaled accuracy and binary accuracy. It is important to recognise that in both the treatment and control groups in both forms of scoring, participants accuracy increased as they continued through the survey.

Figure 4:



Change in participants accuracy from the first 5 emails to the last 10 emails between the treatment and control groups with both the scaled accuracy and binary accuracy.

Discussion

While there is no statistical evidence that our intervention increased accuracy by decreasing overconfidence, there are two primary findings from our study. First, overconfidence is present and does decline more for the treatment than the control group. Second, there is an increase in accuracy meaning participants' ability to detect phishing emails increased as they completed the survey.

First, we will address the role of confidence across the sample. Our goal was to see if overconfidence is a mechanism in people's inability to detect phishing emails. We found that indeed overconfidence is present across the sample as people's confidence ratings at the beginning of the survey were higher than their accuracy scores on the first 5 emails. Thus people overstated their confidence in their ability to properly manage the email inbox. Moreover, if we look at how people adjusted their confidence from the beginning to the end of the survey, 60% of people decreased their confidence suggesting they may have been overconfident to start and then readjusted after going through the survey. So, we can conclude that overconfidence is present and was reduced by completing the survey.

Second, we saw an increase in accuracy from the beginning of the survey to the end for both treatment and control groups. While we would expect to see an increase only with the treatment group due to the intervention, this is not the case. Rather, it may be that taking the survey itself was an experiential learning experience and is the reason accuracy increased for both the treatment and control groups. Our intervention itself may not have been effective, but the survey itself caused an increase in accuracy.

With this, it is important to address a significant limitation which is the within subject design flaw. We wanted to avoid priming people to look out for phishing emails so we did not mention phishing in the briefing of the survey. Although, it is possible participants still became more cautious and aware of the survey as they went through because we still informed participants that there was a point system and that may have been enough to engage their system 2 thinking which is in line with our hypothesis. So, our survey itself may have been the significant intervention which would explain the increased accuracy for the control groups.

Also, apart from the fact that we did not reach our power calculation sample of 506, another limitation is worth acknowledging. Namely that the samples were unbalanced between the treatment and control groups. Through Qualtrics, our survey randomly assigned participants to the treatment or control group. Nonetheless, we see a higher confidence level at the beginning of the survey for the treatment group by 6.5 points and a higher scaled accuracy level with the first five emails for the treatment group by 5 points. We evaluate confidence and accuracy in each group based on the change to help account for these unbalanced samples. Nonetheless, this may have impacted the Mann Whitney U tests to be statistically insignificant. Also, if we assume the participant's increase or decrease in confidence and accuracy are non-linear, then the starting point may impact the participant's susceptibility to change. For example, if an individual in the treatment group started with an accuracy of 90% in the first 5 emails and improved by 10% to 100% accuracy that would appear as quite a significant increase as the knowledge necessary to get all the details may be larger than the knowledge necessary to increase from 50% accurate to 60% accurate. Thus since the treatment group started out with a slightly higher accuracy, the change may be understated relative to the control group.

In addition, we want to acknowledge a concern that our intervention could be making anyone worse off. This means checking that individual's who received the intervention did not become worse at managing the inbox. For example, after receiving the treatment, a participant may become overcautious and rank genuine emails as fake ones. There were 8 participants in the treatment group who chose "Delete and Block" for a higher proportion of the genuine emails from the last 10 emails than the first 5. However, since we do not specify that "Delete and block" should be used for phishing emails, it is possible these individuals just did not want to continue receiving emails from the sender and chose that option as a way to unsubscribe.

Ultimately, our solution does not change significantly. While our research method did not perfectly isolate the treatment from the control, the act of filling out the survey improved the accuracy of both the treatment and control groups. In addition, we saw overconfidence was present in the sample which is in line with Wang, Li & Rao (2016) and Pattinson et al. (2012). Filling out the survey in and of itself is a form of experiential learning and the existence of the point system which participants were informed of may have been enough to make them more conscious of their choices and thoughtful of their failure. This aligns with

the findings of Yoon et. al. (2012) who showed hands-on experiential learning was most effective for improving an individual's cybersecurity compliance. So overall, our results stay in line with our hypothesis that overconfidence is a mechanism at play in one's ability to detect phishing emails and that experiential learning along with confronting people with their failure is an effective method for improving one's ability. One way to improve our solution further is to teach individuals with detailed information on cybersecurity as well after they have experienced failure, as people are better in their uptake after failure compared to a normal situation (Kumaraguru, 2007). In this way, experiential learning and theoretical learning could be combined.

Final Solution and Implementation

Although our analysis did not give any significant results, we do believe that our results in combination with existing literature show that our solution has potential. Therefore, we will now further elaborate on what exactly the solution will entail and how ING could make use of this promising solution by implementing it in real life.

The idea of confronting individuals with their overconfidence and providing them feedback to improve their ability to recognize phishing emails might be fairly difficult to do in a real life situation. The most effective way to do this would be through sending mock attacks to individuals to test their responses. If they indeed fall for the mock attack, they would then be confronted with the fact that this could have lost them personal information and money. As mentioned in the previous section, this confrontation moment could also be used to explain to these individuals how they could be more aware of phishing links in the future. Our results showed that people are in fact overconfident but the intervention did not give a significant result, which could indicate that the intervention was not strong enough. This is why we would advise ING to also use the confrontation moment to show individuals how to recognize a real and a fake URL. Indeed, literature suggests as well, that educating works best after people experienced failure (Wang, Li & Rao, 2016; Kumaraguru, 2007). In this way experiential learning could be combined with some more theoretical learning.

Still, there might be some limitations to this approach. ING would have to be very cautious using this solution. Their customers might not appreciate their "trusted bank" deceiving them by sending them fake phishing attacks; even when this is only for their own benefit. This

could create a negative feeling of customers towards the bank and decrease the chances of customers being open towards the bank helping and supporting them in fighting phishing.

To avoid these negative consequences, we suggest that ING asks people to opt-in to a "phishing assistance program". Without specifying much further what this program entails, ING can let its customers choose whether they want to be helped in fighting their chances of falling for a scheme. ING will then let individuals who opted-in know that for the program to be most effective, they cannot share much more information at this point in time, but that future steps will become clearer eventually. The next step is that ING will send participants of the program monthly "mock" or "simulated" phishing emails. These emails are such that they are very similar in content, look and feel to actual phishing emails, but are in reality sent by ING. Moreover, the phishing link will not forward people to a website where they have to pay or provide information, but to a website ING will create to confront these people with their failure. On this website people will see a text that confronts them with the fact that they just clicked on a phishing link and that next time this could mean they lose real money or important personal information. Additionally, this page will show them the most effective way to recognize a phishing email: judging the URL.

Of course this is not optimal, as there might be selection bias in terms of the people who opt-in; the really overconfident people will feel like they do not need 'help' recognizing phishing. Also, you would want users to be completely unaware of the possibility that ING would send a mock attack. Still we do believe that this program could help a significant amount of people, especially those who know there is a risk but feel like it would not hit them. Moreover, if ING does not feel constrained to send out the fake phishing emails to their customers without consent, the opt-in program is not required at all and ING can simply start sending mock attacks to their customers on a semi-regular basis. This would be the most effective way to execute the solution.

Now we will work out the details from a company perspective, i.e. what resources are needed to make it happen. We think that this solution will actually require very little resources in terms of money and personnel. Of course, money and people will have to be made available to set the program in motion: the fake phishing webpages have to be developed, people have to be made aware of the opt-in program and the emails containing the phishing links must be sent. Also, employees are needed to perform follow-ups and to evaluate the results of this

solution, which we will talk about more later. Still, taking into account the size of the problem, these costs seem fractional.

The most difficult part will be to get as many people to opt-in to the program. One way to do this would be to have new customers opt-out when they want to open a bank account. Using this nudge, people are more likely to stay in the program than they would be to actively sign-up. In addition to this, current customers could be notified through standard channels, such as information letters, social media posts and when meeting with a bank employee. An alternative would be to make customers aware of the possibility when they log-in on their digital banking account or in their app. Again, having the opt-in program is not optimal, but as we are not confident that ING would be willing to send out the simulated phishing emails without permission of the customers, this could be the best option available. The next steps, namely the creation of the confronting webpages and the sending of the emails, should cost very little. Especially when assuming that ING already has in-house developers and automated email systems.

In case the opt-in program would not have sufficient applicants and would thus lose its effectiveness, an alternative could be that individuals are confronted with a short ‘test’ on recognizing phishing emails every 6 months. This test would show whether individuals are actually able to recognize the right emails and could confront them with their results. Customers would for example be required to complete the test before being able to log in to their online banking account. Of course there would then also have to be a button giving the option to “complete later”, so that if people are really in a hurry or have to pay for their groceries with their Apple pay, they do not first have to spend their time on filling out the test. We do believe that there should be a final deadline to complete the test. In terms of costs and capabilities of the implementation of this version of the solution, we think that they are very similar to the ‘mock attack solution’. As such we are convinced both implementation options of our solution are feasible.

Evaluation of the solution

To test the effectiveness of the implementations mentioned above, there are some details that ING should keep in mind. Firstly, ING should continue to send mock emails monthly rather than sending it once. It is expected that users will improve as they are exposed to more of

these attacks, since every time they fall for phishing, their confidence will be reduced and necessary information about how to detect phishing mails will be shown. ING can monitor the progress of users through click rates. They should focus on the difference between previous and latter responses to mock attacks, in other words they should investigate if users are behaving differently across time. If there is a significant difference between different time points, then it can be concluded that the solution is effective enough. To put it differently, clickthrough rate should decrease over time. This assessment could be done every month, and it allows ING to make comparisons at any point in time. Therefore, any progress trend can be easily seen. Furthermore, there is another method that can be used to assess the improvement of users. ING can let individuals know that next time they think they have a fake ING email, they should send it to a certain ING email address. If there is an increase in the number of emails sent to that certain ING response address, made for this purpose, this would confirm people are paying attention and this would also be proof that previous attacks in fact increase users' ability to detect phishing.

Another possible way to understand the effectiveness of the solution is asking people who are in the opt-in program to do a test. When they agree to join the opt-in program, ING can ask them to do a test in which they will face a bunch of phishing emails and try to detect them. At the end of the test, they will get a score. After 6 months from the first test, they will take a very similar test. Meanwhile, ING will keep sending mock attacks during these 6 months. If there would be a significant difference between the initial score and the last score, it would show that there is a considerable increase in awareness and users are much better in detecting phishing thanks to the opt-in program.

Conclusion

People fall victim to phishing attacks far too often. When asked about falling victim to phishing emails, a common statement is "I thought I knew". In this study, we set out to understand the behaviour behind this statement and found an overconfidence bias is present in individual's belief in their ability to manage emails and detect phishing. Luckily, we also found that forms of experiential learning and being faced with failure, can help increase an individual's ability to detect phishing emails. Thus, we propose the solution to ING that they test their clients with mock phishing emails that, when clicked, confront the clients with their failure of falling prey to (mock) phishing. To avoid damaging trust between ING and their

clients, we suggest these mock phishing emails be done through an opt-in program. With initiatives to encourage people to recognize their overconfidence bias and face their failure, hopefully, we can hear less statements of “I thought I knew”.

References

Alba, J. W., & Hutchinson, J. W. (2000). Knowledge calibration: what consumers know and what they think they know. *Journal of Consumer Research*, *27*(2), 123–156.

Anandpara, V., Dingman, A., Jakobsson, M., Liu, D., Roinestad, H., & Dietrich, Sven, spock@cs.stevens.edu, Software Engineering Institute, Carnegie Mellon University, 4500 Fifth Avenue, Pittsburgh, PA, 15213, USA. (2007). Financial cryptography and data security : 11th international conference, fc 2007, and 1st international workshop on usable security, usec 2007, scarborough, trinidad and tobago, february 12-16, 2007. revised selected papers. In *Phishing iq tests measure fear, not ability* (pp. 362–366). essay, Berlin, Heidelberg : Springer Berlin Heidelberg : Springer. https://doi.org/10.1007/978-3-540-77366-5_33

Anderson, J. R. (1993). Rules of the Mind. Lawrence Erlbaum Associates, Inc.

Angner, E. (2006). Economists as experts: overconfidence in theory and practice. *Journal of Economic Methodology*, *13*(1), 1–24. https://doi.org/10.1080/13501780600566271

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behaviour and Human Decision Processes*, *39*(1), 133–144. https://doi.org/10.1016/0749-5978(87)90049-5

Camerer, C. F., & Lovallo, D. (1999). Overconfidence and Excess Entry: An Experimental Approach. *The American Economic Review*, *89*(1), 306–318. https://doi.org/10.1257/aer.89.1.306

Centraal Bureau voor de Statistiek. (2022b, February 28). *2,5 miljoen Nederlanders in 2021 slachtoffer van online criminaliteit*. Centraal Bureau Voor De Statistiek. https://www.cbs.nl/nl-nl/nieuws/2022/09/2-5-miljoen-nederlanders-in-2021-slachtoffer-van-online-criminaliteit

Chuang, W., & Lee, B. (2006). An empirical evaluation of the overconfidence hypothesis. *Journal of Banking and Finance*, *30*(9), 2489–2515. https://doi.org/10.1016/j.jbankfin.2005.08.007

Das, S., Nippert-Eng, C., & Camp, L. J. (2022). Evaluating user susceptibility to phishing attacks. *Information & Computer Security*, *30*(1), 1–18. https://doi.org/10.1108/ics-12-2020-0204

Dhamija, R., Tygar, J. D., & Hearst, M. A. (2006). Why phishing works. *Human Factors in Computing Systems*. https://doi.org/10.1145/1124772.1124861

Fazio, R. H., & Zanna, M. P. (1978). On the predictive validity of attitudes: the roles of direct experience and confidence. *Journal of Personality*, *46*(2), 228–243. https://doi.org/10.1111/j.1467-6494.1978.tb00177.x

Goel, S., Williams, K. L., Triantafilis, J., & Warkentin, M. (2020). Understanding the Role of Incentives in Security Behavior. *Proceedings of the . . . Annual Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2020.519

Hayward, M. L. A., Shepherd, D. A., & Griffin, D. (2006). A hubris theory of entrepreneurship. *Management Science*, *52*(2), 160.

Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, *4*(1), 5–28. https://doi.org/10.1007/bf00057884

Hewitt, B., & White, G. L. (2022). Optimistic Bias and Exposure Affect Security Incidents on Home Computer. *Journal of Computer Information Systems*. https://doi.org/10.1080/08874417.2019.1697860

Hirshleifer, D., & Luo, G. Q. (2001). On the survival of overconfident traders in a competitive securities market. *Journal of Financial Markets*, *4*(1), 73–84. https://doi.org/10.1016/s1386-4181(00)00014-8

Hong, J. (2012). The state of phishing attacks. *Communications of the ACM*, *55*(1), 74–81. https://doi.org/10.1145/2063176.2063197

Hong, K. W., Kelley, C. M., Tembe, R., Murphy-Hill, E., & Mayhorn, C. B. (2013). Keeping up with the joneses : assessing phishing susceptibility in an email task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 1012–1016. https://doi.org/10.1177/1541931213571226

Kahneman, D. (2011). *Thinking, Fast and Slow*. Doubleday Canada.

Kassin, S. M., Rigby, S., & Castillo, S. (1991). The accuracy^confidence correlation in eyewitness testimony: Limits and extensions of the retrospective self-awareness effect. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.61.5.698

Kumaraguru, P., Rhee, Y. G., Sheng, S., Hasan, S., Acquisti, A., Cranor, L. F., & Hong, J. (2007). Getting users to pay attention to anti-phishing education. *Proceedings of the Anti-Phishing Working Groups 2nd Annual ECrime Researchers Summit*. https://doi.org/10.1145/1299015.1299022

Lei, W., Hu, S., & Hsu, C. (2022). Unveiling the Process of Phishing Precautions Taking: The Moderating Role of Optimism Bias.

Libby, R., & Rennekamp, K. (2011). Self-serving attribution bias, overconfidence, and the issuance of management forecasts. Journal of Accounting Research, 50(1), 197-231.

Metzger, M. J., & Suh, J. J. (2017). Comparative optimism about privacy risks on facebook. *Journal of Communication*, *67*(2), 203–232. https://doi.org/10.1111/jcom.12290

Moore, D. A., & Healy, P. M. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295x.115.2.502

Moores, T. T., & Chang, J. C.-J. (2009). Self-efficacy, overconfidence, and the negative effect on subsequent performance: a field study. *Information & Management*, *46*(2), 69–76. https://doi.org/10.1016/j.im.2008.11.006

Parsons, K., McCormac, A., Butavicius, M. A., & Ferguson, L. (2010b). Human Factors and Information Security: Individual, Culture and Security Environment. *Defence Science and Technology Organisation Edinburgh (Australia) Command Control Communications and Intelligence DIV*.

Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., & Jerram, C. (2013). Phishing for the truth: A scenario-based experiment of users' behavioural response to emails. *In Security and Privacy Protection in Information Processing Systems: 28th IFIP TC 11 International Conference, SEC 2013, Auckland, New Zealand, July 8-10, 2013. Proceedings 28 (pp. 366-378).* Springer Berlin Heidelberg.

Pattinson, M., Jerram, C., Parsons, K., McCormac, A., & Butavicius, M. (2012). Why do some people manage phishing emails better than others? *Information Management & Computer Security*, *20*(1), 18–28. https://doi.org/10.1108/09685221211219173

Posner, M. I., Snyder, C. R., & Solso, R. (2004). Attention and cognitive control. *Cognitive psychology: Key readings*, *205*, 55-85.

Russo, J. E., & Schoemaker, P. J. (1992). Managing overconfidence. *Sloan management review*, *33*(2), 7-17.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207.

Simon, H. A. (1986). Rationality in psychology and economics. Journal of Business, S209-S224.

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), 945

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behaviour and Human Decision Processes*, *42*(3), 271–283. https://doi.org/10.1016/0749-5978(88)90001-5

Symantec, I. N. C. (2014). Internet security threat report. Mountain View, CA:[sn], 44**.**

Tang, F., Hess, T. J., Valacich, J. S., & Sweeney, J. T. (2014). The effects of visualization and interactivity on calibration in financial decision-making. *Behavioral Research in Accounting*, *26*(1), 25–58. https://doi.org/10.2308/bria-50589

The Brussels Times. (2022, December 30). *Over four in ten Belgians victims of phishing last year*.https://www.brusselstimes.com/344394/over-four-in-ten-belgians-victims-of-phishing-last-year

Wang, J., Li, Y., & Rao, H. R. (2016). Overconfidence in phishing email detection. *Journal of the Association for Information Systems*, *17*(11), 759–783. https://doi.org/10.17705/1jais.00442

Weinstein, N. D., Marcus, S. E., & Moser, R. P. (2005). Smokers' unrealistic optimism about their risk. *Tobacco Control*, *14*(1), 55.

Yoon, C., Hwang, J.-W., & Kim, R. (2012). Exploring factors that influence students' behaviors in information security. *Journal of Information Systems Education*, *23*(4), 407–415.

Appendix

Appendix 1: Survey prompt and initial question on confidence

Dear,

Thank you for participating in our research, we greatly appreciate it.

In the following you will be shown 15 screenshots of emails in the inbox of Jack Johnson, which he received from various companies. We are interested in understanding how you would manage the inbox of Jack Johnson.

Please try your best to organise the emails in this mailbox, by suggesting an appropriate action for each email. You will be scored based on how you answer. The possible responses are 1. Flag for follow up, 2. Leave in inbox, 3. Delete, 4. Delete and block.

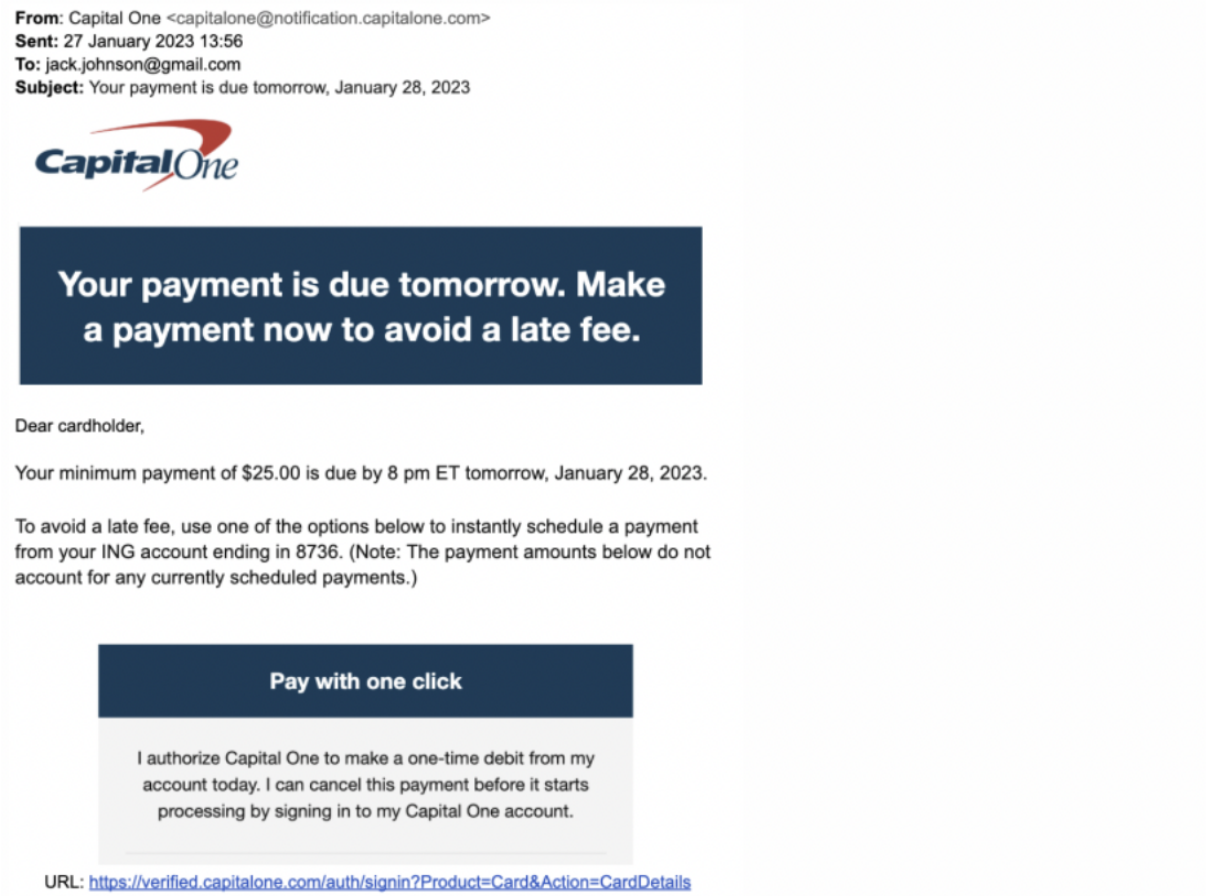
If you give the most appropriate response you can receive a maximum amount of 2 points per email. The minimum amount of points is 0.5. The scoring is divided uniformly among questions.

Before starting we would like to ask you to fill out some additional questions.

On a scale of 1-10, how confident are you that you can choose the correct action for each email?

Appendix 2: Email example

Appendix 2.1: Genuine email used in the survey



From: Capital One <capitalone@notification.capitalone.com>
Sent: 27 January 2023 13:56
To: jack.johnson@gmail.com
Subject: Your payment is due tomorrow, January 28, 2023

Capital One

Your payment is due tomorrow. Make a payment now to avoid a late fee.

Dear cardholder,

Your minimum payment of $25.00 is due by 8 pm ET tomorrow, January 28, 2023.

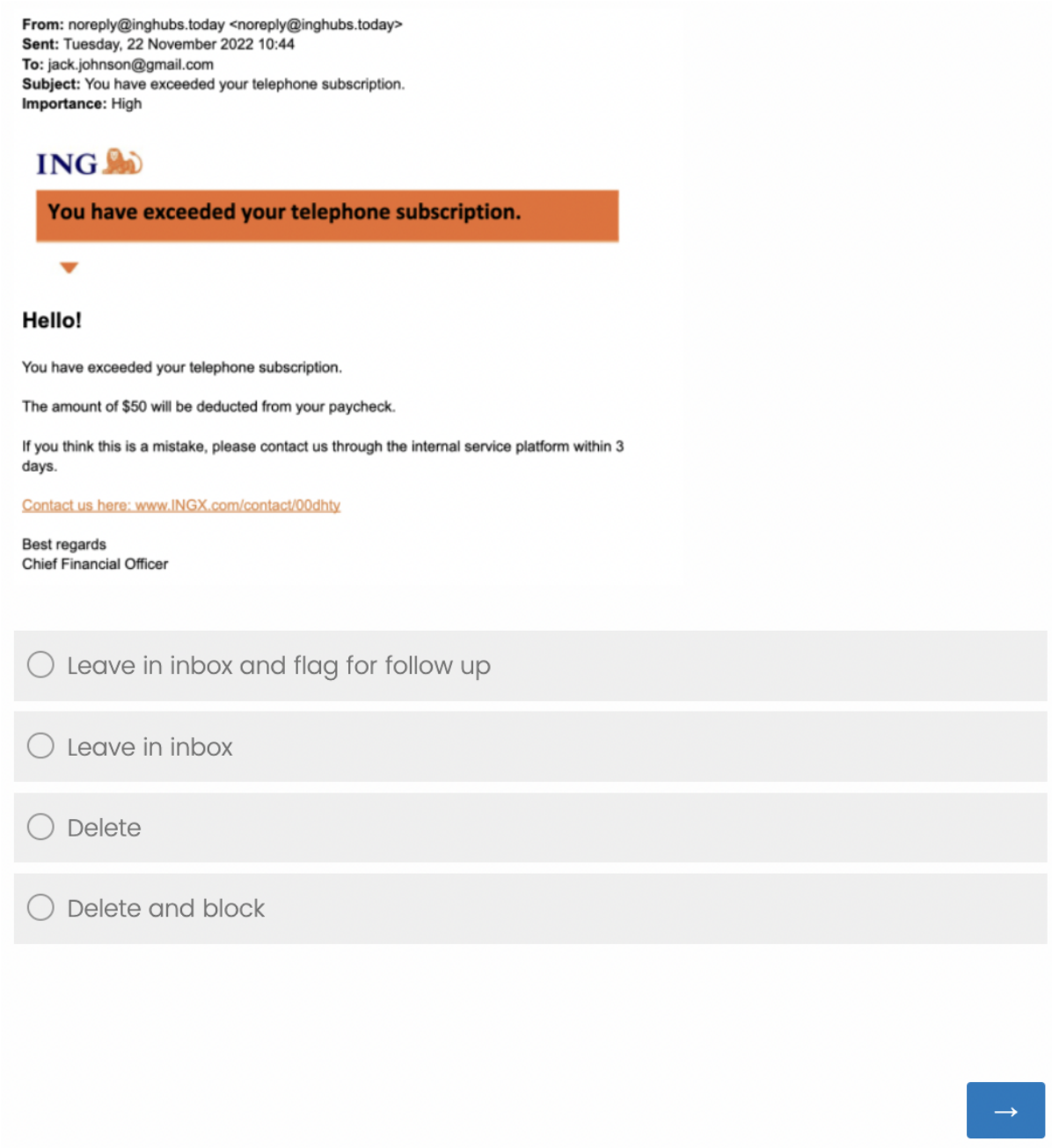
To avoid a late fee, use one of the options below to instantly schedule a payment from your ING account ending in 8736. (Note: The payment amounts below do not account for any currently scheduled payments.)

Pay with one click

I authorize Capital One to make a one-time debit from my account today. I can cancel this payment before it starts processing by signing in to my Capital One account.

URL: https://verified.capitalone.com/auth/signin?Product=Card&Action=CardDetails

Appendix 2.2: Phishing emails used in the survey



From: noreply@inghubs.today <noreply@inghubs.today>
Sent: Tuesday, 22 November 2022 10:44
To: jack.johnson@gmail.com
Subject: You have exceeded your telephone subscription.
Importance: High

ING

You have exceeded your telephone subscription.

Hello!

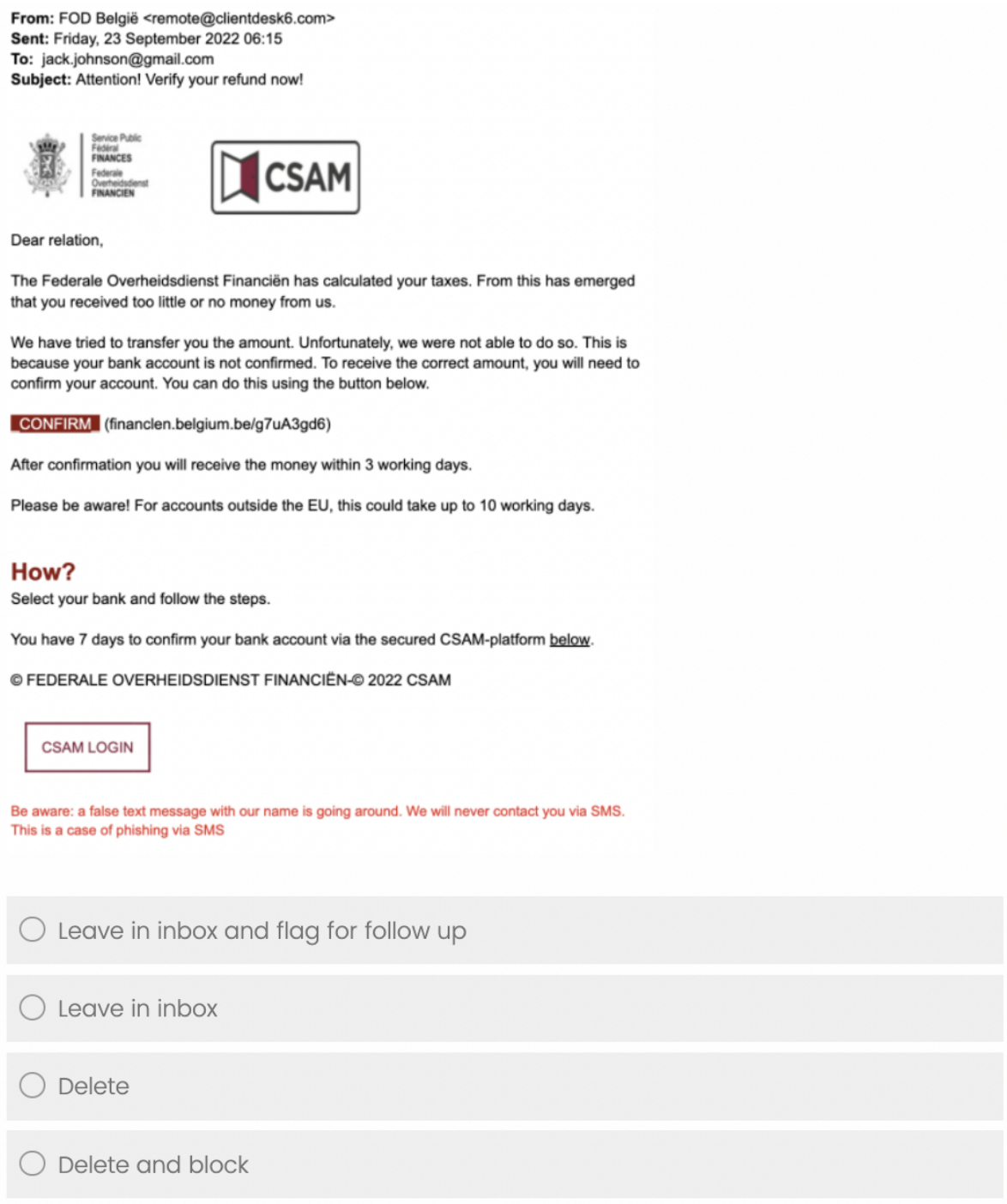
You have exceeded your telephone subscription.

The amount of $50 will be deducted from your paycheck.

If you think this is a mistake, please contact us through the internal service platform within 3 days.

Contact us here: www.INGX.com/contact/00dhty

Best regards
Chief Financial Officer



Appendix 3: Intervention for the treatment group in the survey

Message:

You were X% correct in providing the most appropriate way to manage the previous 5 emails. You believed you would be Y% confident. IF(confidence > correct, This means you were overconfident).

Example 1 (confidence = 80% and accuracy =70%):
You were 70% correct in providing the most appropriate way to manage the previous 5 emails. You believed you would be 80% confident. This means you were overconfident.

Example 2 (confidence = 70% and accuracy =75%):
You were 75% correct in providing the most appropriate way to manage the previous 5 emails. You believed you would be 70% confident.

Appendix 4: Debriefing statement

Thank you for participating. Your answers will be used anonymously. The goal of our research is to find out whether people are overconfident about recognizing phishing emails and whether their responses improve when they are confronted with a poor performance. Please be aware of phishing emails every time you open your own inbox!

Appendix 5: Power Calculation

Based on previous research on the matter (Pattinson et.al, 2012) we believed the mean score to be a little above 50% of the total score. As participants could get a score of 20 points in total, we set the mean for our power calculation on 13. Furthermore, we were interested in finding a 1 point difference and set the standard deviation to be 4, again based on the literature. Indeed after collecting the data and performing our analysis the standard deviation turned out to be a little below 4. To perform the calculation we used the standard settings for the Type-I and Type-II errors, which are 0.05 and 0.2 respectively.