MASTER

Investigating the Effect of Phishing Sophistication on Phishing Reporting

Kersten, Leon

Award date:
2021

Investigating the Effect of Phishing Sophistication on Phishing Reporting

Leon Kersten
l.kersten.1@student.tue.nl
Eindhoven University of Technology

ABSTRACT

The reporting of phishing emails is crucial for organizations to detect phishing attacks that are getting ever more sophisticated. Despite extensive research on how the sophistication of a phishing attack affects detection rates, there is little to no research about the relationship between the sophistication of a phishing attack and the associated reporting rate. In this work, we perform a controlled experiment with 446 subjects with eight experiment conditions to evaluate how the reporting rate is linked to the sophistication of a phishing email and its detection rate. Each experiment condition is a phishing email composed of different levels of sophistication based on four factors: 'Technical', 'Contextual', 'Language and Tone' and 'Layout'. Our results show that the reporting rate decreases as the sophistication of the attack increases without any evidence of inflection points and that around half of the people has an intention to report an email given that one detects the phishing mail. However, the group intending to report an email is not a subset of the group detecting a phishing email, suggesting that reporting is still a concept misunderstood by many.

CCS CONCEPTS

• **Security and privacy → Phishing**.

KEYWORDS

phishing, reporting, controlled experiment

1 INTRODUCTION

Phishing attacks continues to become more sophisticated than ever. Cybercriminals gather more information about the victim, to craft highly targeted and personalized emails to deceive one into handing over their credentials or executing a payment towards the attacker [12]. While the sophistication of phishing attacks increases, there seems to be no sign a decrease of the frequency of such attacks. About 70% of cyberattacks use some sort of social engineering within their process [12] and 43% of data breaches in the last few years involved phishing techniques [26]. Despite the high frequency of phishing attacks, countermeasures against highly sophisticated attacks are limited [1]. More specifically, existing technical solutions such as network monitoring and filtering or protective technologies do not work as effectively against more sophisticated attacks [1], creating a need to rely on humans to detect phishing emails [4]. However, individuals being able to detect phishing emails would only solve part of the problem. Individuals need to notify relevant departments within their organization about the existence of phishing campaigns using some reporting mechanism such that the relevant department knows about the attack, and can act accordingly. Despite the importance of the concept of reporting in the context of phishing, the change in reporting behavior with respect to the sophistication of an attack is relatively unexplored. Additionally, from past work it is unclear what the relationship is between an individual detecting a phishing email and deciding to report the email.

This work investigates the relationship between the level of sophistication of a phishing attack and the reporting rate of the phishing email used in such attack. For this, we devise a framework to define sophistication in quantifiable means. Additionally, compare the group of individuals who decide to report a phishing email with individuals who detect the same email, to get a clearer understanding of the 'reporting group'. To achieve these two aims, we perform a controlled experiment with 446 subjects through Amazon Mechnical Turk. In the controlled experiment, the respondents read one of the many possible emails with different levels of sophistication, and are asked to answer questions about reporting and detection of that email.

The remainder of this paper is structured as follows. Section 2 presents background information on what defines a sophistication of a phishing email and on the concept of reporting and discusses related work. Then, the research questions of this paper are presented in Section 3 and Section 4 explains the methodology used to answer those research questions. Section 5 provides the results of our experiment and Section 6 discusses our findings. Finally, Section 7 provides conclusions.

2 BACKGROUND AND RELATED WORK

Sophistication in the context of phishing has different meanings across different literature. Sophistication of a phishing email can refer to how much money the attacker has spent, how well the email could avoid blacklisting [14], or even whether the phishing attack makes use of an malware or not [10]. In this work, we are interested in how well the email is tailored to the target and therefore in this paper we use the terms believability and level of sophistication interchangeably to refer to the same concept. A multitude of aspects that influence the detection of phishing emails have been substantially studied in previous studies. These aspects can make an email more or less believable for the target and increase or decrease the level of sophistication in phishing attacks.

There is a plethora of different 'cues' that increase the suspicion of a phishing email for the victim and thus decrease the believability of a phishing email [20, 22]. Commonly seen 'cues' in less sophisticated phishing emails include spelling mistakes or bad grammar[15, 22, 27] and spoofed links [9, 22]. In addition, usage of spoofed sender names and domain are common practice for phishers [15] and email recipients who pay greater attention to sender

, ,
.

Table 1: The Four Factors of Phishing Sophistication

Factors	Definition
Technical (T) [15, 27]	The technical specifications of the email such as the sender name and domain.
Contextual (C) [5, 8, 22]	The phishing mail's semantics w.r.t the pretext of the victim.
Language and Tone (Lg) [15, 22, 27]	The use of appropriate language and tone with respect to the context of the victim.
Layout (Ly) [22, 29]	The degree in which the visual aspects closely resembles what the victim would expect from a non-phishing mail in their context.

addresses are more likely to detect a phishing email [27]. This suggests that the utilization of spoofed sender names may increase the sophistication of a phishing attack. Moreover, visual cues such as usage of logos or conventional formatting of emails influence the believability of a phishing email [22]. A realistic and accurate visual design increases the user's trust towards the email and increases its persuasiveness [29]. In an extreme case, a study showed that two similar phishing emails, one with an authentic design and one without had a response rate of 47% and 19% respectively [29].

Additionally to the deployment of 'cues' in phishing emails, previous work include studies [5, 25, 30] implementing the cognitive vulnerabilities identified by Cialdini, namely Authority, Liking, Scarcity, Consistency, Social Proof and Reciprocity [6]. The six cognitive vulnerabilities can be used to increase the persuasiveness of a phishing email [30]. However the effect of implementing these cognitive vulnerabilities seem to differ greatly between studies. Additionally, different cognitive vulnerabilities show vastly different results with respect to the click rate of a phishing email [5]. Moreover, the direction to which the believability changes seem to be crucially dependent on the context and the environment of the potential victim. As an example, in a study implementing Authority via adding contact information had opposite effects on two samples working for different institutions [5].

On this same line, *Greene et al.* further stress the importance of context and premise alignment for phishing mails [8]. Their work shows that an individual's professional context influences the way one looks at a phishing email. More specifically, if the premise of the phishing email aligns with an individual's context, the individual is more likely to focus on the aspects that make the phishing email believable and ignore suspicious cues that may indicate a phishing attack. Conversely, if the email does not fit the context of the recipient, the recipient is more likely to focus on the suspicious cues. Since premise alignment determines how much effect the existence of cues or usage of cognitive vulnerabilities have, it can potentially be one of if not the most important factor in determining the level of sophistication of a phishing email. Overall, the literature identifies four factors that can influence the believability of a phishing email; these are summarized in Table 1.

Next to phishing *detection*, phishing *reporting* is a relatively new mechanism studied in the literature. Reporting phishing emails in an organizational context may allow IT personnel of the organization to detect an ongoing phishing attack earlier [1, 11]. This may allow the organization to warn other employees and deploy any additional countermeasures in a timely manner [11]. Such timely response is crucial in mitigating the impacts of large scale phishing campaigns considering that most of the victims will appear in the first few hours [5] from the start of the campaign. Unfortunately, reporting rates seem to be drastically lower than detection rates, potentially because there is a plethora of incentives for recipients to not report phishing emails such as but not limited to: fear of the negative consequences of misreporting [11], distrust in the capabilities of the IT department [11] and an unclear reporting mechanism [21]. Conversely, cases have been reported where a subject had the intention to report an email to verify if the email is legitimate without necessary having the belief that the email was a phishing email [4]. Despite the substantial research in phishing mails using click rate or response rate as metrics, there is still little work where reporting rate is used as a metric while reporting is of high importance for organizations targeted by a phishing campaign. Importantly, the link between the features of the attack in terms of its believability and the chances that an email is reported is not yet evaluated in the extant literature.

Research Gap The overall literature review suggests that no clear relationship between reporting behavior and the level of sophistication of phishing emails has been well defined in past literature. Understanding such relationship, can drive future research to develop more effective measures for phishing response and containment by better employing reporting mechanisms for phishing emails.

3 PROBLEM STATEMENT AND RESEARCH QUESTIONS

Problem statement. From this point onward, we refer to sophistication and believability of a phishing email, as believability only. The literature suggests that the relationship between user behaviour and the believability of a phishing email is multifaceted and not straightforward; for example, Burda et al. [5] showed that additional contextual information added to an attack may be counterproductive for the attacker depending on the environment in which the attack is deployed. Considering phishing *reporting* specifically, to date it is unclear what the relationship is between attack features in terms of believability of the attack, and a user's decision to report it or not. For example, a group of users may only report if they feel the need to verify the legitimacy of an email. Therefore, such groups may not report 'obvious' phishing emails as they may believe that such phishing emails are so obvious that there is no need to personally verify those emails to check its legitimacy. If this is the case, one would observe a complex relationship between the (increasing) believability of a phishing mail, and the likelihood of reporting: for example, the number of people reporting an email may decline slower or even rise as the believability of a phishing email increases, because a group feels the need to verify the legitimacy of the email at that level of believability.

Research question. Considering the above problem statement and the research gap, we pose the following research question:

To what extent does the believability of a phishing email affect the detection and reporting rates of that email?

To answer this question, we formulate the following four sub research questions:

- **RQ1**: What is the effect of the email features listed in Table 1 on the believability of an email?
- **RQ2**: What is the relation between the believability of a phishing email and its detection rate?
- **RQ3**: What is the relation between the believability of a phishing email and its reporting rate?
- **RQ4**: How does detection of a phishing email by the target affect the reporting rate?

4 METHODOLOGY

Overview of method. To investigate the correlation between the believability of an email and its reporting rate, we design and run a controlled experiment in which subjects answer questions by means of a questionnaire about their willingness and rationale to report or not report specific (phishing) emails. Each questionnaire shows the user an email that the subject must read before providing their answers to the questions. Subjects are randomly assigned to an experiment condition in the controlled experiment. We devise eight experiment conditions by means of a (2^{4-1}) fractional factorial design, each implementing different combinations of techniques at increasing the believability of the email, based on the literature review summarized in Table 1. We evaluate the believability of each experiment condition using control questions. The control questions measures how well each technique of Table 1 have been implemented. The exact technique we use to evaluate the believability can be found in Section 4.5.

4.1 Experimental subjects

We conducted the experiment on Amazon Mechanical Turk [3] with a sample size of 446 subjects. Within Amazon Mechanical Turk, we have only selected U.S residents as subjects for two reasons. First and foremost, the content of the crafted phishing email is tailored towards U.S residents (see 4.2), making the phishing email potentially less effective when presented outside that context. Secondly, U.S respondents for Amazon Mechanical Turk tend to be quite representative of the U.S population for the purposes of security and privacy questionnaires [18], while it may not necessary be the case that global respondents are representative of the global population.

Additionally, we imposed a constraint on the subjects that they need more than 1000 HITs approved and a HIT approval rate of 98% or greater. HIT is a term used by Amazon Mechnical Turk to refer to any single, self-contained virtual task that a MTurk worker can do on the platform. We imposed such constraints such that we only recruit subjects who have a reputation of being trustworthy within MTurk as they have shown that they have completed tasks on the platform previously. Furthermore, by only recruiting individuals with an experience on MTurk, we increase the likelihood that our subjects will be familiar with the interface and process of filling in a questionnaire linked to MTurk.

After the subjects filled in the questionnaire, any results from questionnaires that we could not trace the identity to a MTurk account has been removed from the data set. This ensures that double entries from a single person are not counted, as it would add a bias towards our data. Additionally, all subjects who completed the survey within 90 seconds have been rejected and removed from the experiment, as we deemed it impossible to fill in the questionnaire seriously within that time frame [1]. Last but not least, any subjects who answered unintelligibly, in the open questions (see Section 4.4) has been removed from the data set as well. We took lenient approach towards what is considered intelligible, and only replies consisting of pseudo random strings, or words without any relation to the questionnaire ordered in an incomprehensible manner are removed from the data set.

4.2 Experiment Conditions

For the purposes of this experiment, we consider the treatment factors outlined in Table 1 as booleans, i.e. a feature which the email may or may not have. Each feature can have a level of *high*, meaning that it is implemented, or *low*, meaning that it is not implemented. For example, an email can have the value *high* for technical while *low* for context, layout, and language. In this case the phishing email would mimic the technical aspects of the sender it is attempting to imitate, while its use of language and the visual aspects do not represent what an original email would have looked like in the specific context. Furthermore, these factors are also assumed to be independent from each other. In other words, implementing one of the factors in a phishing email should not influence the impact of another implemented (or not implemented) factor in that email. We discuss limitations introduced by this assumption in Section 6.1 and 6.2.

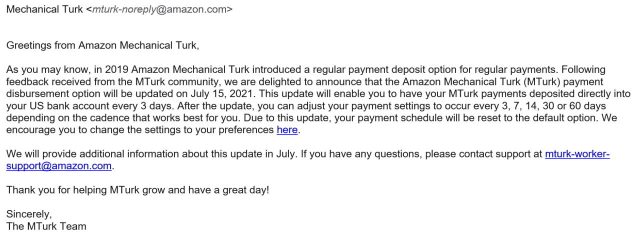
From this section onwards, phishing emails used in this experiment are referred to with combinations of abbreviations introduced in Table 1. The presence of an abbreviation implies that the corresponding factor's value is *high* and consequentially the absence of an abbreviation implies that the value of that factor is *low*. As an example, an email which has values *high* for technical (T) and layout (Lg) is referred to as **TLg**. An email which has value *low* for all factors is referred to as **None**.

Implementing all possible variations with four binary variables would require the creation of sixteen experimental conditions, which would prove difficult to manage both for the experiment design and implementation, and for the recruitment of enough subjects to obtain statistically valid insights on the investigated phenomenon. To mitigate this, we employ a 2^{4-1} fractional factorial design [13] such that only eight experiment conditions are needed. The experiment conditions consists of all phishing emails with two factors having the value *high* and two factors having the value *low* (six different emails in total), an email where all factors have the value *low* and an email where all factors have the value *high*. Our experiment conditions are therefore: **None**, **TC**, **TLg**, **TLy**, **CLg**, **CLy**, **LgLy** and **TCLgLy**.

4.3 Phishing pretext

As our experiment is conducted on Amazon Mechnical Turk, we use a previously sent email from Amazon Mechnical Turk [23] to model a baseline phishing email. This ensures that our respondents are somewhat familiar with the context of the email. The baseline email can be seen in Figure 1. We modified the email to update dates such that those corresponded with the time frame in which

[1] The expected completion time of the survey is 5 minutes.



Mechanical Turk <*mturk-noreply*@amazon.com>

Greetings from Amazon Mechanical Turk,

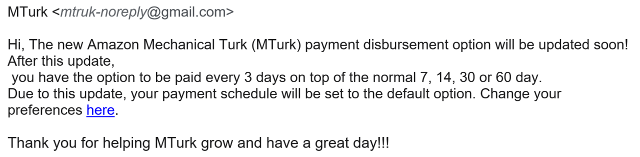
As you may know, in 2019 Amazon Mechanical Turk introduced a regular payment deposit option for regular payments. Following feedback received from the MTurk community, we are delighted to announce that the Amazon Mechanical Turk (MTurk) payment disbursement option will be updated on July 15, 2021. This update will enable you to have your MTurk payments deposited directly into your US bank account every 3 days. After the update, you can adjust your payment settings to occur every 3, 7, 14, 30 or 60 days depending on the cadence that works best for you. Due to this update, your payment schedule will be reset to the default option. We encourage you to change the settings to your preferences here.

We will provide additional information about this update in July. If you have any questions, please contact support at mturk-worker-support@amazon.com.

Thank you for helping MTurk grow and have a great day!

Sincerely,
The MTurk Team

Figure 1: The baseline phishing email used in the experiment (TCLgLy)



MTurk <*mtruk-noreply*@gmail.com>

Hi, The new Amazon Mechanical Turk (MTurk) payment disbursement option will be updated soon!
After this update,
you have the option to be paid every 3 days on top of the normal 7, 14, 30 or 60 day.
Due to this update, your payment schedule will be set to the default option. Change your preferences here.

Thank you for helping MTurk grow and have a great day!!!

Figure 2: The email after applying all treatments to remove the presence of all four factors (None)

the experiment was conducted and to introduce a hyperlink in the email such that the email was asking the user to take a specific action. The latter is important to mimic a phishing email which often asks the respondent to do an action rather than only reading the email [28].

We use this baseline email as the most ‘believable’ one (i.e., **TCLgLy**) as it is the least modified version from the original Amazon Mechanical Turk email. To reproduce the other experimental conditions, we will then remove each feature accordingly moving the implementation of each factor from *high* (the baseline email) to *low* (the treatment condition). An overview of modifications is provided in Table 2.

A phishing email where all these treatments are implemented (**None**) can be seen in Figure 2. Additionally, each separate treatment is reported in Table 7 in the appendix.

4.4 User Questionnaire

After being randomly assigned to an experimental condition (and therefore an email), each subject is asked a set of questions. There are three purposes to this questionnaire: 1) to collect data to measure the detection and reporting rate of the experiment conditions; 2) to measure how well the four relevant factors have been implemented in the experiment conditions; 3) to explore the rationale of why individuals may or may not report or detect an email. Before receiving answering the set of questions, the subjects were notified what ’reporting’ refers to in the context of phishing attacks.

The exact wording of the questions used in the questionnaire is reported below, together with its purpose, in brackets: *Data* denotes the first purpose of data collection, *Control* denotes the second purpose of checking the implementation of a factor and *Rationale* denotes the third purpose of gathering rationale for reporting and detection behavior. Furthermore, conditional questions are marked with a letter after the question number and only appear depending on the response in the question before. The requirements to trigger a conditional question is specified in italics after the question. The three conditional questions are open or semi-open (multiple choice with an option to answer the question as an open question) such that we do not restrict the respondents in their rationale related to their reporting behavior while all other questions are multiple choice questions.

1 Email users may choose to either report or **not report** for various reasons. How likely are you to report this email to Amazon? (*Data*)

2 Conditional Questions

- 2*a* What are the main reasons why you are not likely to report this email? (*Rationale*)
  (*Condition: respondent answers Neutral, Unlikely or Extremely Unlikely for question* 1.)
- 2*b* What are the main reasons why you are likely to report this email? (*Rationale*)
  (*Condition: respondent answers Likely or Extremely likely for question* 1.)

3 If you received this email in your inbox, how likely would you be to believe that the email is a legitimate communication from Amazon Mechanical Turk? (*Data*)

4 Conditional Question.

- 4*a* What aspects of the email make you think that Amazon Mechanical Turk is not likely to have sent it to you? (*Rationale*)
  (*Condition: respondent answers Neutral, Unlikely or Extremely Unlikely for question* 3.)

5 If you received this email in your inbox, how likely would you be to believe that the sender’s email address is a legitimate email address of Amazon Mechanical Turk? (*Control*)

6 If you received this email in your inbox, how likely would you be to believe that the message the email communicates is credibly from Amazon Mechanical Turk? (*Control*)

7 How well does the used language and tone of the email above match any previous email you received from Amazon Mechanical Turk? (*Control*)

8 The layout of an email refers to the visual elements of the email such as logos, and line and paragraph spacing. How well does the layout of the email above match any previous email you received from Amazon Mechanical Turk? (*Control*)

We reduced ambiguity in the interpretation of the questions by the subjects by running two question review rounds: in the first round, relevant literature [7, 17] was consulted to minimize the common ‘pitfalls’ in questionnaire wording. In addition to ambiguity, some questions such as question 1 may also be subject to desirability bias [17] as reporting a potential phishing email may be considered a socially desirable action.[2] In the second review round, the preliminary questions were discussed with multiple individuals, including a native English speaker to reduce the ambiguity and streamline the questions wording. After two rounds of discussion, the wording of the questions was finalized.

[2] To mitigate this, two countermeasures were employed in question 1: firstly, it starts with a statement indicating that both reporting and not reporting are acceptable [17] . Secondly, as suggested in [17, 19] ’not report’ is shown in bold to balance the question and makes the respondent less hesitant to indicate that they would not report an email.

Table 2: Modifications to the baseline email to implement the experimental conditions

Factor	Modification to the baseline email (**TCLgLy**)
Technical	The sender address is modified from mturk-noreply@amazon.com to mtruk-noreply@gmail.com, a fake email address which uses the Gmail domain, a domain that is relatively easy to obtain and a misspelled sender name. Additionally, the display name changes from “Mechanical Turk” to “MTurk”, a commonly used shorthand notation of Amazon Mechanical Turk.
Contextual	Any mentions that the email is announcing a new feature is removed, and existing features are announced in the email. Furthermore, any indication of time with respect to when the new feature should be available is removed to make the email more generic and less specific to the context of the person.
Language	Several words and phrases commonly used in emails by Amazon Mechanical Turk is replaced with more informal alternatives. As an example, the iconic phrase “Greetings from Amazon Mechanical Turk” present in most emails sent by Amazon Mechanical Turk is replaced with the generic informal greeting: “Hi”. Additionally, some exclamation marks are added to sentences where normally no exclamation would be used in the context of Amazon Mechanical Turk.
Layout	Incoherent paragraph spacing are implemented and pseudo-random line spacing have been added between words within sentences. Furthermore, the font sizes of some words are increased from 10.5 to 11.

Response scales and interpretation. We use a five-point Likert scale for questions 1, 3, 5, 6, 7 and 8. To minimize ambiguity in the responses, we rely on [17, 24] to choose our wording for the Likert scales. Considering survey ethics [17], additional to the five-point Likert scales, subjects have the option to answer ’Don’t Know’. All responses to questions with Likert Scales are interpreted in a binary fashion. The binary interpretation of the Likert scale is done as otherwise there might be insufficient sample size for each response to analyze the data in a meaningful manner considering the limited size of the total number of subjects. When a subject answers ’Likely’ or ’Extremely Likely’ to a question (or its relevant alternative wording if a different wording is used for the Likert scale in that specific question), the response is considered to be a positive response. For example for question 1, if a respondent answers that they are ’Likely’ to report an email, we consider the respondent to have an intent to report the email presented in the questionnaire. When any other answer than ’Likely’, ’Extremely Likely’ or ’Don’t Know’ is given (or again any of the relevant alternatives), we consider the response to be a negative response. Responses which answered ’Don’t Know’ in one or multiple questions are unused for parts of the data analysis where the response of that question is considered crucial.

4.5 Believability Model

To approximate the relative level of believability while using the four factors introduced in Section 2 and summarized in Table 1, we observe that the believability should increase the more the email is *believable*, i.e. the more the reader considers the information provided therein as credibly from the source (in our case, Amazon MTurk). We therefore estimate how believable a phishing email may be based on the credibility expressed by the MTurk users of our implementation of each of the four factors. In other words, we use the results from the control questions to estimate the believability of a given email. To achieve this, for each factor we divide the experiment conditions into two groups. One with value *high* for a specific factor and the other group with the value *low*. Then for each group we count the number of subjects who have given a positive response or a negative response (as defined above) for the relevant control question. Based on these counts we estimate the odds ratio (OR) of a positive response, which we identify as an indication that the treatment increases the believability of the email. Consider the following hypothetical scenario. In total 10 respondents have answered ’Likely’ or ’Extremely Likely’ for question 5 in the questionnaire while reading an email with T on value *high*. Meanwhile 5 respondents answered ’Neutral’, ’Unlikely’ or ’Extremely Unlikely’ for the same question. From the group of people who read an email with T on value *low*, 6 answered ’Likely’ or ’Extremely Likely’ and 12 respondents answered otherwise. In this case, $OR_T = (10/5)/(6/12) = 4$. Using the estimate for all ORs for each factor, we will then estimate a ’believability index’ for a phishing email by linearly combining all factors identified in Table 1:

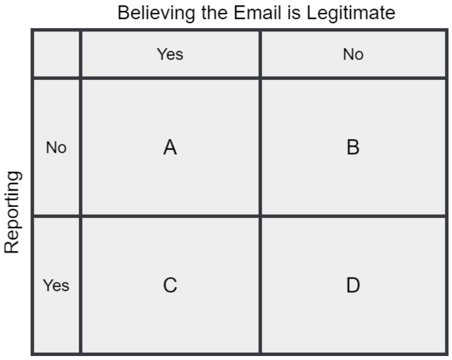
$$B = OR_T T + OR_C C + OR_{Lg} Lg + OR_{Ly} Ly$$

where B denotes the believably index (hereafter, believability) of a phishing mail, OR_x denotes the OR for factor $x \in T, C, Lg, Ly$; each factor is identified as the corresponding binary variable for taking value of 1 when present (i.e., *high*) and 0 otherwise (i.e., *low*).

4.6 Evaluating Effects on Reporting

To isolate the effects per factor given the users’ belief that the email is legitimage, we obtain ORs by counting occurrences of four different outcomes, as summarized in the figure below: As there are eight experiment conditions, we have eight 2x2 matrix with those counts for the relevant responses.

To isolate the effects we rely on our fractional factorial experiment design; for each factor we separate the eight count matrices into two groups: a group of four matrices with counts of experiment conditions with that specific factor *high* and another with *low*. Then, we calculate the respective mean of the number of subjects that had the same experiment outcome for each group. We estimate effect significance on the OR estimation by using Fisher’s Exact Test. We consider significance level $\alpha < 0.05$ as statistically significant.



	Yes	No
No	A	B
Yes	C	D

Figure 3: An example on how the counts where organized per experiment condition. A, B, C, D are all counts and thus positive integers.

4.7 Ethical Aspects

This research was executed under ethical approval from our institution's ERB under file ERB2020MCS13.

5 RESULTS

5.1 Evaluation of email believability (RQ1)

Table 3 shows the odds ratio estimate of the probability of a positive response to the control questions (questions 5 to 8). An odds ratio greater than one indicates that the probability of a positive response to the corresponding control question increases when the corresponding factor is changed from *low* to *high*. Conversely, an odds ratio less than one indicates that the probability of a positive response decreases when the corresponding factor is changed from *low* to *high*. Firstly, we notice that all factors have an odds ratio above 1 and that all factors except C have a p-value smaller than 0.05, indicating that there is a significant difference in the distribution of positive responses across the experiment conditions. This suggests that the factors T, Lg and Ly are implemented in such a way that the implementation noticeably increases the probability that a responded would find the respective treatment implementation credible, and therefore increase the believability of the email. The treatment implementation is especially credible for factor T, where the odds ratio is the highest with 3.98. The implementation is less noticeable for Lg and Ly where the odds ratio is lower than that of Technical with 1.88 and 1.79 respectively.

Under the assumptions described in Section 4.5, we can then identify the effect sizes of each factor on the overall 'believability' of an email, namely:

$$B = 3.98T + 1.08C + 1.88bLg + 1.79Ly$$

Table 3: OR Estimate with p-value for the Control Questions

Factor	Odds Ratio	Confidence Interval	p-value
Technical	3.98	[2.62, 6.11]	< 0.001
Contextual	1.08	[0.73, 1.61]	0.37
Language & Tone	1.88	[1.21, 2.94]	< 0.001
Layout	1.79	[1.19, 2.69]	< 0.001

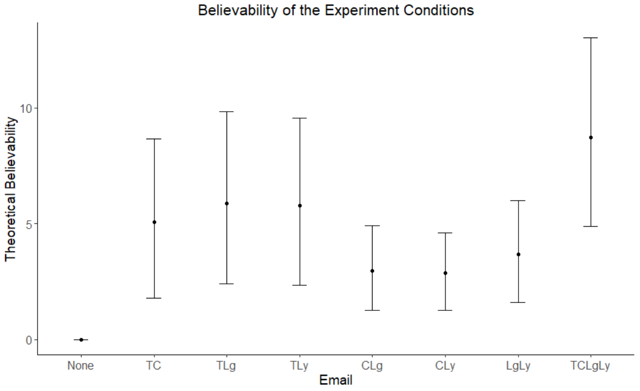


Figure 4: Theoretical Believability for each Possible Factor Combination

Table 4: Response Count for Believing the Legitimacy of the Email
The largest count for each row are shown in bold. Additionally, subjects who responded 'Don't Know' for question 3 are not displayed in this table.

Email	Subjects	Legitimate(%)	Not Legitimate(%)
None	61	30(49.2)	**31(50.8)**
TC	59	**42(71.2)**	17(28.8)
TLg	59	**44(74.6)**	15(25.4)
TLy	73	**55(75.3)**	18(24.7)
CLg	53	26(49.1)	**27(50.9)**
CLy	48	22(45.8)	**26(54.2)**
LgLy	39	17(43.6)	**22(56.4)**
TCLgLy	50	**42(84.0)**	8(16.0)
Total	442	**278(62.9)**	164(37.1)

Figure 4 shows the theoretical believably for our eight experiment conditions calculated from the equation given in Section 4.5. The confidence intervals have been estimated by assuming that the logarithm of the odds ratio shown in Table 3 is normally distributed. Based on this assumed distribution, we sampled 10000 values for each factor. For this we used the empirically estimated odds ratio for each factor as the mean and we calculated the standard deviation based on the confidence interval of the estimated odds ratio for each factor. Using the samples, we calculate the confidence interval from the respective linear combinations needed to calculate the believability for each email. Figure 4 shows that all emails with T having the value *high* have a higher theoretical believability compared to any emails where T has the value *low*. This means that the within this model the factor T has a significant contribution in defining the believability of a phishing email.

5.2 Analysis of Detection Rates (RQ2)

Table 4 shows the response count for question 3 for each experiment condition, or in other words the number of respondents who believe their email to be, respectively, a legitimate or not legitimate

Table 5: Odds Ratio of Reporting and Believing Legitimacy Count for Each Factor
p-values under 0.05 are shown in bold.

	5a: Believing Legitmacy (L)			5b: Reporting (R)			5c: $R\|L$			5d: $R\|\neg L$		
Factor	OR	CI	p-value	OR	CI	p-value	OR	CI	p-value	OR	CI	p-value
Technical	0.29	[0.12, 0.69]	**0.002**	2.77	[1.14, 7.01]	**0.011**	2.70	[0.67, 11.37]	0.117	1.42	[0.34, 6.12]	0.749
Contextual	1.02	[0.44, 2.35]	0.597	1.10	[0.46, 2.66]	0.490	0.87	[0.23, 3.33]	1.000	1.42	[0.37, 5.52]	0.763
Language and Tone	0.90	[0.39, 2.09]	0.474	1.38	[0.57, 3.40]	0.279	2.23	[0.54, 11.03]	0.237	0.90	[0.23, 3.53]	1.000
Layout	0.88	[0.38, 2.02]	0.442	1.10	[0.46, 2.66]	0.490	0.80	[0.20, 3.18]	0.767	1.43	[0.36, 5.81]	0.757

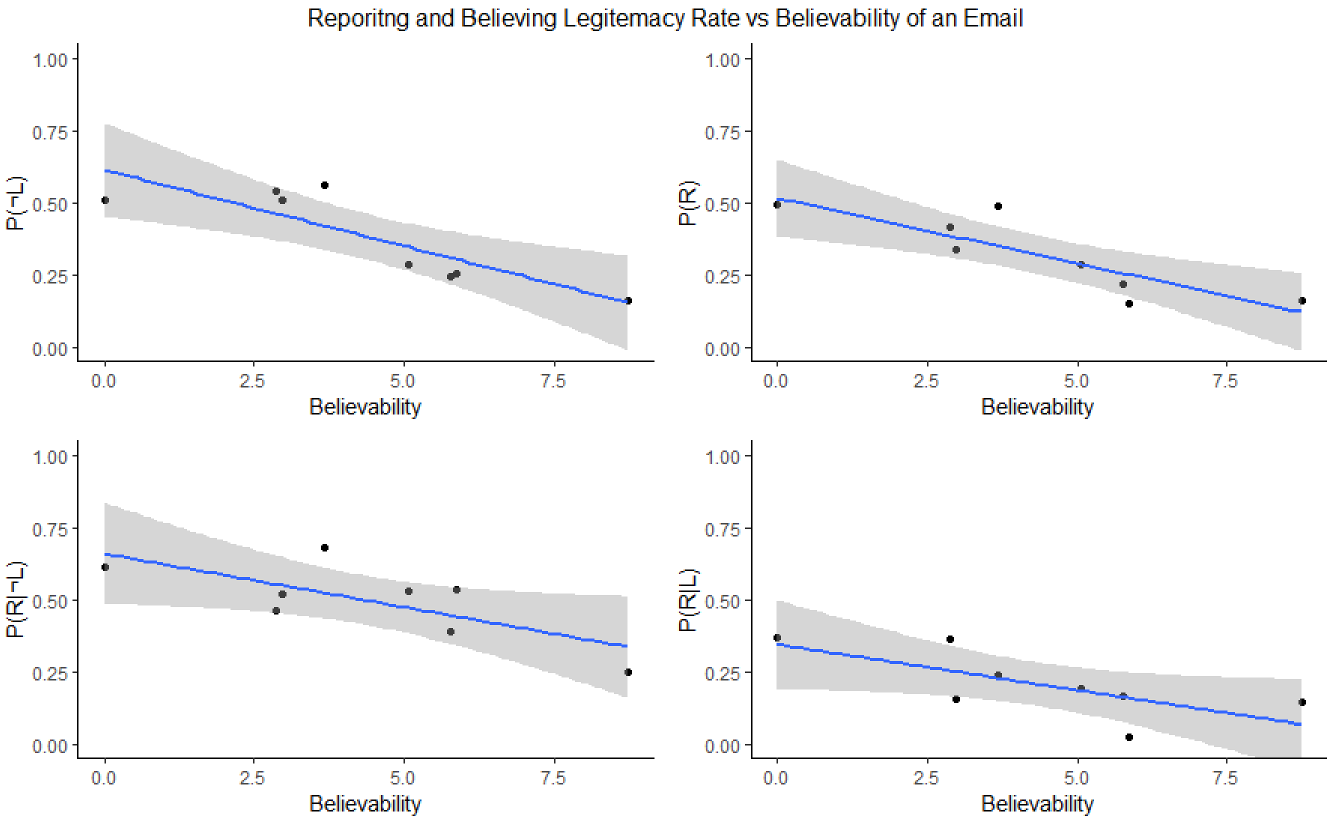


Figure 5: Relationship between {P(R),P(¬ L),P(R|¬ L),P(R|L)} and the Believability of an Email

email from MTurk. We consider respondents who believe that their email is not a legitimate email from MTurk to have detected the email as a phishing email. Conversely, respondents who believe the email to be legitimate are considered to not have detected the phishing email. The total sample size from this sub section onwards is 442 instead of 446, as 4 subjects answered 'Don't Know' for question 3. Therefore, it is impossible to know whethere these four respondents detected or did not detect the phishing email. In total 278 out of the 442 or 62.9% of all respondents did not detect the experiment condition as a phishing mail. By contrast, 37.1% of all respondents detected the email as a phishing mail. The email with the lowest detection rate is **`TCLgLy`** where 16% believed that the email is a legitimate email from MTurk. Thus, the email with the highest ranked believiability attained the lowest detection rate. The experiment condition with the highest detection rate (i.e highest number of people believing the email to be illegitimate), is **`LgLy`** where 56.4% of the respondents believed that the email is not a legitimate email from MTurk. Surprisingly, despite the implementation of factor `C` being unnoticeable (see Table 3), **`CLg`** and **`CLy`** both have lower detection rates than **`LgLy`** with 50.9% and 54.2% respectively. Furthermore, the detection rate of the least believable email according our model, namely **`None`** is also lower than that of **`LgLy`** with 50.8%. This indicates that simply having more factors on value *high* may not necessary decrease the detection rate, even if the phishing email is more believable according to our model.

Table 5 shows an overview of the results for all ORs for each factor computed as described in Section 4.6. Of relevance here is the first column (5a) which shows the OR and its significance for detection. Table 5a shows that the factor `T` is the only factor with a sufficiently low p-value to conclude that the detection rate significantly changes depending on whether the factor has value *high* or *low*. As the OR for `T` is 0.29, if `T` changes from *high* to *low*, the number of people who believe the email to be legitimate decreases by four times. In other words, the detection rate increases by four times. For all other factors, there is insufficient evidence to suggest that the factor influences the detection rate.

Figure 5 shows several graphs where different rates such as detection and reporting rate are plotted versus the theoretical believability for each experiment condition. The top-left section of Figure 5 specifically shows the graph where the detection rate ($P(\neg L)$) is plotted. From this, we find that in general less people detect a phishing email as the believability of the mail increases. Therefore, the detection rate decreases as phishing emails become more believable.

5.3 Analysis of Reporting Intentions (RQ3)

6 shows an overview of the results relating to the reporting rate. In total, 137 out of the 442 or 31% of all respondents showed an intention to report the email given in the questionnaire (see Table 6a). The email with the highest reporting rate is **None** where almost half of the respondents (49.2%) showed an intent to report that email. Meanwhile, the email with the lowest reporting rate is **TLg** with a reporting rate of 15.3% although **TCLgLy** has only a marginally higher reporting rate (16.0%). Similar to detection, only factor `T` has a sufficiently low p-value to conclude that the detection rate significantly changes depending on whether the factor has value *high* or *low*. As the OR for `T` is 2.77 (see Table 5a), if `T` changes from *high* to *low* the reporting rate increases almost three times. However, as the confidence level is quite wide (1.14 to 7.01), the magnitude of the coefficient of the increase in reporting rate can differ significantly.

The top-right graph in Figure 5 shows the reporting rate plotted versus the believably of each experiment condition. The reporting rate ($P(R)$) decreases as the believability of a phishing email increases. Since all but two points is within the confidence interval of the linear line of fit, it is possible that the relationship between the reporting rate and the believability of a phishing email is linear.

5.4 The Influence of Detection on the Reporting Rate (RQ 4)

From the respondents who detected the mail shown in the questionnaire as a phishing email, about half (52.4%, see Table 6b) of the respondents showed an intent to report. Interestingly, taking aside **TCLgLy**, the reporting rate given that the respondent did detect the email as a phishing mail does not seem to differ greatly, taking ranges from 38.9% to 68.2%. The fact that **TCLgLy** has a much lower reporting rate when respondents do detect the email may be attributed to the low sample size (8 respondents) for respondents who read **TCLgLy** and believed that this email was not legitimate. The fact that the range of reporting rate when detecting the email as a phishing mail is not very wide, suggests that the reporting rate may not differ across different levels of believability when an individual detects the mail. This suggestion is further strengthened when considering each factor that compromises the experiment conditions separately. Table 5b shows that no factor has a sufficiently low p-value to conclude that reporting rate given that one detects the email as a phishing mail changes when a factor is changed from value *high* to *low*. However the bottom-left graph in Figure 5 displays a decreasing trend of the reporting rate given detection ($P(R|\neg L)$) as the believability of the email increases. Yet, if the outlier **TCLgLy** is removed from the data set, $P(R|\neg L)$ does not have a decreasing trend, strengthening our initial suggestion in this section.

Moreover, 47.6% (see Table 6b) of the respondents who detected the email as a phishing mail, showed no intention to report. The most common reason (covering 41.0% of responses) among this group for having no intention to report is that they found reporting too 'Time Consuming'. Additionally, 28.2% indicated that they do not know how to report and hence would not report emails even when believing that an email is a phishing email.

Interestingly, in total 18.3% (see Table 6c) of the respondents has the intention to report an email despite believing it is a legitimate email from MTurk. For some of our experiment conditions, such as **None** and **CLy**, the proportion of individuals who report while believing the email is legitimate is higher than a third (36.7% and 36.4% respectively). Additionally, the bottom-left graph in Figure 5 shows that the reporting rate of respondents who believe that their email is legitimate ($P(R|L)$ decreases slightly as believiability increases although with a much lower gradient than that of $P(R)$ and ($P\neg L$). However, Table 5c shows that no factor has a sufficiently low p-value to conclude that reporting rate given that the respondent did not detect the email changes when a factor is changed from value *high* to *low*.

6 DISCUSSION

Characterization of the reporting group. Our findings show that in contrast to real life reporting rates shown in past literature [11, 16], the percentage of people having the intent to report a phishing email is much higher. The difference in reporting rates between our study and in practice may be due to the fact that actually reporting an email in practice takes more effort than reporting an email by answering 'Likely' or 'Extremely Likely' on a questionnaire. Therefore, many individuals may show an intent to report on the questionnaire, yet not report when a similar incident happens in real life. Moreover, despite the reporting rates in our research being higher than other studies, the group of people having an intent to report a phishing email is still quite small. Even our phishing mail with the highest reporting rate did not achieve a reporting rate of more than 50%. Furthermore, if we look at the reporting rate throughout all of our experiment conditions not even a third has intention to report an email while 37.1% detected the email as a phishing email.

However, unlike intuition may suggest, the set of people wishing to report their email is not a subset of the people who detected their email as a phishing mail. For example, 18.3% of the individuals who

Table 6: Response Count
The percentages shown under columns 'Legitimate' and 'Not Legitimate' are calculated from the total number of respondents within that row who believed their email to be legitimate or illegitimate respectively, not the total respondents within that row. Furthermore, the largest count for each row and categories are shown in bold.

		6a: Overall (%)		6b: Not Legitimate (%)		6c: Legitimate (%)	
Email	Subjects	Report	Not Report	Report	Not Report	Report	Not Report
None	61	30(49.2)	**31(50.8)**	**19(61.3)**	12(38.7)	11(36.7)	**19(63.3)**
TC	59	17(28.8)	**42(71.2)**	**9(52.9)**	8(47.1)	8(19.0)	**34(81.0)**
TLg	59	9(15.3)	**50(84.7)**	**8(53.3)**	7(46.7)	1(2.3)	**43(97.7)**
TLy	73	16(21.9)	**57(78.1)**	7(38.9)	**11(61.1)**	9(16.4)	**46(83.6)**
CLg	53	18(34.0)	**35(66.0)**	**14(51.9)**	13(48.1)	4(15.4)	**22(84.6)**
CLy	48	20(41.7)	**28(58.3)**	12(46.2)	**14(53.8)**	8(36.4)	**14(63.6)**
LgLy	39	19(48.7)	**20(51.3)**	**15(68.2)**	7(31.8)	4(23.5)	**13(76.5)**
TCLgLy	50	8(16.0)	**42(84.0)**	2(25.0)	**6(75.0)**	6(14.3)	**36(85.7)**
Total	442	137(31.0)	**305(69.0)**	**86(52.4)**	78(47.6)	51(18.3)	**227(81.7)**

believed their email to be legitimate showed an intention to report the email and almost half (47.6%) of the individuals who detected their mail as a phishing attack showed no intention to report the email. This suggests that the set of people who report emails and the set of people who detect phishing emails, are two distinct sets. These sets do have an overlap but are in no means equal or have the relation where one is a subset of another.

The rationale given to report an email while believing that that email is legitimate illustrates that reporting as a concept is often misunderstood by individuals. Some tend to see reporting as a behavior to indicate that they are happy by the contents of the email. For example, one subject who showed an intention to report while believing the email is legitimate, gave the following rationale for reporting: *"It sounds so great, because i can get the payment as per my wish from 3 days to 60 days"*. Meanwhile, some people may report an email just to check whether it is legitimate. For example a respondent who answered that they are likely to report the email and are likely to believe the email is legitimate gave the following rationale for reporting: *"just to get a clarification and see if it is official"*. In practice such behavior may lead to a substantial amount of false alarms, hindering the detection of actual phishing campaigns. Therefore, our study suggests there may be a need to raise awareness what the purpose of reporting is to individuals such that the reporting mechanism can be used effectively to detect targeted phishing campaigns. Additionally, there is a need make the large group who do manage to detect the phishing email yet show no intention to report, to change their behavior to report such emails such that false alarms by individuals who misunderstand reporting are diluted and to increase the chance of earlier detection of phishing campaigns. Perhaps this can be achieved by raising more awareness about the impact that reporting has or to provide incentives such that individual who manage to detect phishing emails will be more likely to report. If such incentives are provided however, organizations have to implement it smartly in such a way that individuals do not report all emails to claim their incentives. Future research can focus on these possible methodologies to increase the reporting rate without significantly increasing the false positive rate of reported emails.

Effects of the believability of a phishing email to the reporting rate. Our results show that the believability of a phishing email, can affect the reporting rate of an email. Since, the reporting rate decreases linearly as phishing emails get more believable, the means for the relevant departments to detect such phishing campaigns become harder as well. This is especially problematic as technical tools also tend to become less effective in detecting phishing emails as the believability increases [16].

Furthermore, our results shows that not all factors contribute equally to the overall decrease in reporting rate. Namely, the factor Technical has a significant contribution in the decrease, while we found no evidence of Language and Tone, Context [3] and Layout affecting the reporting rate of an email. This suggests that in practice phishing campaigns utilizing spoofed email addresses will have substantially lower reporting rates, making such emails much harder to detect by the relevant IT personnel. If this is the case a need arises for individuals to be able to detect phishing emails even if the email address overlaps with the imitated sender.

As phishing emails with spoofed email addresses are much more effective than other phishing emails from an attacker perspective, future research could focus more in the area of spoofed phishing emails. More specifically, future research could investigate effective countermeasures to attackers utilizing spoofed email addresses, or methodologies to effectively make individuals better at detecting and reporting such mails.

6.1 Limitations

There are numerous limitations within our study. Firstly, our sample is based on Amazon Mechnical Turk workers (MTurkers) which reside in the United States. Although literature suggests that U.S

[3] Note that the implementation of the Context was not significant as opposed to the implementation of Language and Tone and Layout which was detected by the subjects.

MTurkers are an acceptable representation of the U.S population relevant to other forms of questionnaires [18], the MTurker population in the US generally tends to be slightly younger and higher educated than the average U.S citizen [17, 18]. Although it may be the case that these people interact the most with phishing emails, it was not in the scope of our study to investigate this. Therefore, it is a possibility that our sample size does not represent the average U.S citizen that occasionally interacts with a targeted phishing campaign. However, what is even more problematic is that our sample only accounts for U.S residents. As different cultures may have different ways approaching cybersecurity problems and interacting with potential phishing emails [2], our results may not apply for countries other than the U.S.

Secondly, since our study utilized a questionnaire rather than a real life phishing campaigns, we only measure whether the respondents has an intention to report an email but not whether they would actually report an email. We have not verified whether respondents who showed an intent to report actually report emails outside this questionnaire and hence we cannot guarantee that this study effectively measures reporting rate in practice. Although, we attempted to reduce biases in the questionnaire which could disproportionately skew the responses from real life behavior, as explained in Section 4.4, we did not measure whether such attempts had an effect in reducing the potential biases for the respondents. Furthermore, there may be a plethora of unaccounted biases which cause the reporting behavior of individuals to be different than in the questionnaire.

Thirdly, the factor Contextual has been incorrectly implemented in our experiment and thus the results for the control questions did not show significance for factor Contextual. Therefore, we cannot strongly conclude that the effect of factor Contextual is insignificant or not w.r.t to the decrease of the reporting rate. The wrongful implementation of this factor is due to the fact that the *'new feature'* that we wanted to introduce while impersonating MTurk in our emails, was already introduced recently in reality without our knowledge. Therefore, there was little difference in the factor Contexual being *low* and *high*. It may very well be possible that if the factor Contextual has been correctly implemented that this factor has an influence on the reporting rate. Therefore, within this study we can only strongly conclude what effects Technical, Language and Tone and Layout has on the reporting rate as a whole.

Finally, there is an underlying assumption in our experiment that the effects of the four factors (T, C, Lg and Ly) are independent from each other. In other words, we assume that when one factor changes from *high* to *low* or vice versa, that the effects of the other factors stay equal. However, this is a strong assumption which may not necessary be true. For example, it could be the case that individuals notice 'cues' like bad grammar and wrong spelling easier, when a line break is added after the wrongfully spelled word. In this case, changing Ly to *low* could change the strength of factor Lg without any modification to the implementation of Lg. Possibilities of such effects could be researched in future works to understand the extent of this limitation in our work.

6.2 Threats to Validity

Construct Validity. Our experiment connects the concept of believability to four factors, T, C, Lg and Ly. However, we do not test whether believability from a target's perspective is tied to those four factors, and hence it is a possibility that actual believability is not as strongly related to these factors unlike what this paper suggests. Furthermore, even if these four factors are the core of the concept of believably, it is possible that the treatments applied in the experiment do not correspond with those four factors. For example, as discussed in Section 6.1, the treatment related to C did not correspond with the actual intent of implementing factor C. Additionally, we cannot examine if our subjects have read the email in the survey before answering their questions. To mitigate this, we rejected all respondents who answered the questionnaire under 90 seconds, meaning that all our respondents put in sufficient time to answer the questionnaire.

Internal Validity. We have assumed the independence of the four factors in our model to estimate the believability of an email. However, it is unlikely that these four factors are completely independent from each other and thus this assumption is too strong. This means that the estimated believability of our experiment conditions in our results do not exactly represent true believability of that email, as dependencies of factors may make an email even more or less believable. Additionally, when collecting data we split responses of questions using five-point Likert scale in two categories as explained in Section 4.2. Here the response 'neutral' is categorized as a negative response, because we interpreted that being neutral towards a likelihood of undertaking an action (e.g reporting) would mean that the respondent is more likely to be indecisive. Since being indecisive more often results in an action not being undertaken, the response 'neutral 'is counted as a negative response. However, respondents may understand 'neutral' as a positive reponse which challenges our internal validity. Therefore, some of the data in our experiment categorized as a negative response may be a positive response if the respondent interprets neutral in such a way.

External Validity. The sample consists of experienced MTurkers, who are based in the U.S. Considering that MTurkers on average are slightly younger, and higher educated than the average person, our results may not generalize to older people or people who do not use technology on a regular basis. Additionally, as our sample is U.S based, our results may not generalize to the rest of the world where interpretations of emails may be different [2], and thus where our treatments in our experiment may not apply in the same magnitude as our experiment.

7 CONCLUSION

Our research shows that the relationship between the reporting rate and the believability of a phishing email is a decreasing relationship without any evidence of inflection points. This finding suggests that it is possible to assume lower reporting rates for more sophisticated phishing attacks in countermeasures utilizing reporting mechanisms. Additionally, our results highlight that reporting rates are much higher when individuals are able to detect the phishing email. This suggests that increasing the ability of individuals to detect phishing mails contribute to a degree to higher reporting rates. However, our results also show that reporting is still often misunderstood, and thus one may not be able to blindly follow the reporting of individuals until the concept of reporting phishing email is better understood by society. Furthermore, our results highlight the power of spoofed emails addresses in impacting the reporting rate but also the detection rate of the email. The effect of spoofed email addresses is significantly greater than effects from other factors may have such as layout and language. This suggests that further research may be needed to focus on and effectively counter spoofed phishing emails.

ACKNOWLEDGMENTS

This work is supervised by Luca Allodi and co-supervised by Nicola Zannone and Pavlo Burda.

REFERENCES

[1] Luca Allodi, Tzouliano Chotza, Ekaterina Panina, and Nicola Zannone. 2020. The Need for New Antiphishing Measures Against Spear-Phishing Attacks. *IEEE Security Privacy* 18, 2 (2020), 23–34. https://doi.org/10.1109/MSEC.2019.2940952

[2] Ibrahim M. Alseadoon, Rabie A. Ramadan, and Ahmed Y. Khedr. 2017. Cultural impact on Users' Ability to protect themselves against Phishing websites. *IJCSNS International Journal of Computer Science and Network Security* 17, 11 (2017). http://paper.ijcsns.org/07_book/201711/20171101.pdf

[3] Amazon. [n.d.]. *Amazon Mechanical Turk.* Retrieved September 18, 2021 from https://www.mturk.com/

[4] Pavlo Burda, Luca Allodi, and Nicola Zannone. 2020. Don't Forget the Human: a Crowdsourced Approach to Automate Response and Containment Against Spear Phishing Attacks. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS PW).* 471–476. https://doi.org/10.1109/EuroSPW51379.2020.00069

[5] Pavlo Burda, Tzouliano Chotza, Luca Allodi, and Nicola Zannone. 2020. Testing the Effectiveness of Tailored Phishing Techniques in Industry and Academia: a Field Experiment. In *ARES '20: Proceedings of the 15th International Conference on Availability, Reliability and Security.* https://doi.org/10.1145/3407023.3409178

[6] Robert B. Cialdini. 1984. *Influence: The Psychology of Persuasion.* Harper Business.

[7] Robert Colosi. 2005. Negatively Worded Questions Cause Respondent Confusion. *ASA Section on Survey Research Methods* (2005). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.556.243&rep=rep1&type=pdf

[8] Kristen K. Greene, Michelle P. Steves, Mary F. Theofanos, and Jennifer Kostick. 2018. User Context: An Explanatory Variable in Phishing Susceptibility. In *Workshop on Usable Security (USEC).* USEC, San Diego, CA. https://dx.doi.org/10.14722/usec.2018.23016

[9] Markus Jakobsson. 2007. The Human Factor in Phishing. (2007). http://citeseerx.ist.psu.edu/viewdoc/similar?doi=10.1.1.68.8721&type=cc

[10] Engin Kirda and Christopher Kruegel. 2005. Protecting users against phishing attacks with AntiPhish. In *29th Annual International Computer Software and Applications Conference (COMPSAC'05)*, Vol. 1. 517–524. https://doi.org/10.1109/COMPSAC.2005.126

[11] Youngsun Kwak, Seyoung Lee, Amanda Damiano, and Arun Vishwanath. 2020. Why do users not report spear phishing emails? *Telematics and Informatics* 48 (2020). https://doi.org/10.1016/j.tele.2020.101343

[12] Microsoft. 2020. *Microsoft Digital Defense Report.* Technical Report.

[13] NIST and SEMATECH. [n.d.]. *Engineering Statistics Handbook.* Retrieved September 23, 2021 from https://www.itl.nist.gov/div898/handbook/pri/section3/pri3347.htm

[14] Adam Oest, Yeganeh Safaei, Penghui Zhang, Brad Wardman, Kevin Tyers, Yan Shoshitaishvili, and Adam Doup. 2020. PhishTime: Continuous Longitudinal Measurement of the Effectiveness of Anti-phishing Blacklists. In *29th USENIX Security Symposium (USENIX Security 20).* USENIX Association, 379–396. https://www.usenix.org/conference/usenixsecurity20/presentation/oest-phishtime

[15] Kathryn Parsons, Malcolm Pattinson Marcus Butavicius, Agata McCormac, Dragana Calic, and Cate Jerram. 2015. Do Users Focus on the Correct Cues to Differentiate Between Phishing and Genuine Emails? *Cues of Phishing Emails* (May 2015). https://arxiv.org/ftp/arxiv/papers/1605/1605.04717.pdf

[16] Proofpoint. 2021. *2021 State of the Phish.* Technical Report.

[17] Elissa M. Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L. Mazurek. 2017. *A Summary of Survey Methodology Best Practices for Security and Privacy Researchers.* Technical Report. University of Maryland.

[18] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. 2019. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *2019 IEEE Symposium on Security and Privacy (SP).* 1326–1343. https://doi.org/10.1109/SP.2019.00014

[19] Eric M. Shaeffer, Jon A. Krosnick, Gary E. Langer, and Daniel M. Merkle. 2005. Comparing the Quality of Data Obtained by Minimally Balanced and Fully Balanced Attitude Questions. *Public Opinion Quarterly* 69, 3 (Jan. 2005), 417–428. https://doi.org/10.1093/poq/nfi028

[20] Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. 2020. What makes phishing emails hard for humans to detect? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 64, 1 (2020), 431–435. https://doi.org/10.1177/1071181320641097

[21] Nathalie Stembert, Arne Padmos, Mortaza S. Bargh, Sunil Choenni, and Frans Jansen. 2015. A Study of Preventing Email (Spear) Phishing by Enabling Human Intelligence. In *2015 European Intelligence and Security Informatics Conference.* 113–120. https://doi.org/10.1109/EISIC.2015.38

[22] Michelle Steves, Kristen Greene, and Mary Theofanos. 2020. Categorizing human phishing difficulty: a Phish Scale. *Journal of Cybersecurity* 6, 1 (Sept. 2020). https://doi.org/10.1093/cybsec/tyaa009

[23] u/roads30. 2019. *Amazon Payments is changing things. Just got this email.* Retrieved September 15, 2021 from https://www.reddit.com/r/mturk/comments/dt6743/amazon_payments_is_changing_things_just_got_this/

[24] Wade M. Vagias. 2006. *Likert-Type Scale Response Anchors.* Technical Report. Clemson University.

[25] Amber van der Heijden and Luca Allodi. 2019. Cognitive Triaging of Phishing Attacks. In *Proceedings of the 28th USENIX Security Symposium.* USENIX, Santa Clara.

[26] Verizon. 2021. *Data Breach Investigations Report.* Technical Report.

[27] Jingguo Wang, Tejaswini Herath, Rui Chen, Arun Vishwanath, and Raghav H. Rao. 2012. Research Article Phishing Susceptibility: An Investigation Into the Processing of a Targeted Spear Phishing Email. *IEEE Transactions on Professional Communication* 55, 4 (Sept. 2012), 345–362. https://doi.org/10.1109/TPC.2012.2208392

[28] Emma J. Williams and Adam N. Joinson. 2020. Developing a measure of information seeking about phishing. *Journal of Cybersecurity* 6, 1 (Feb. 2020), 417–428. https://doi.org/10.1093/cybsec/tyaa001

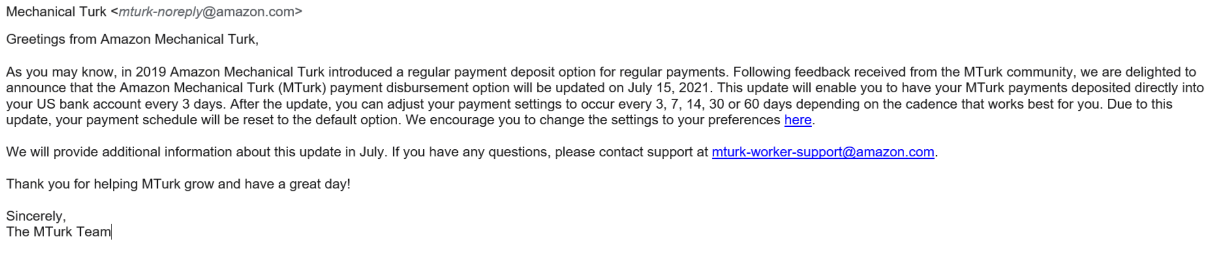
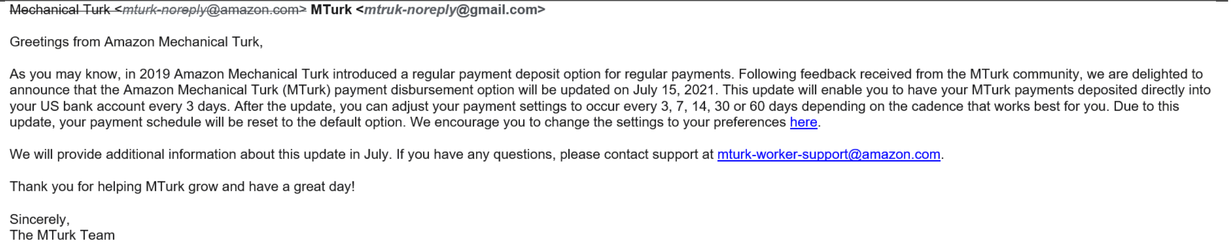
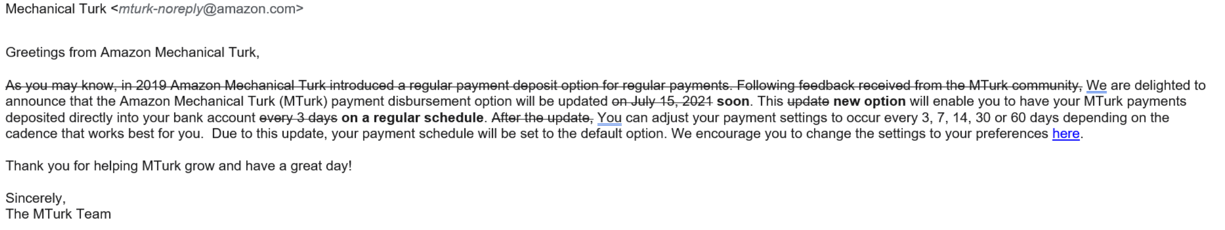
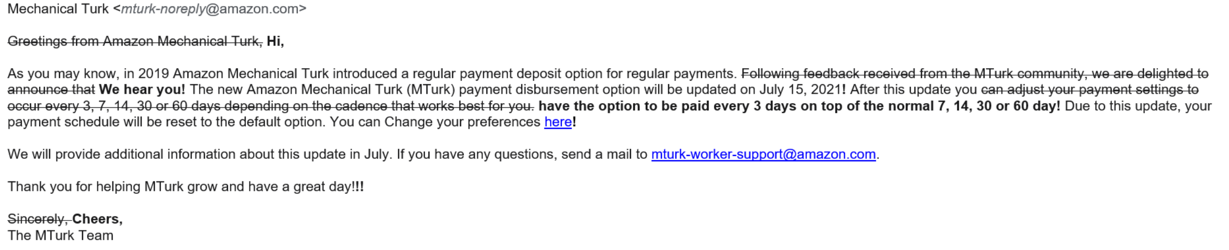
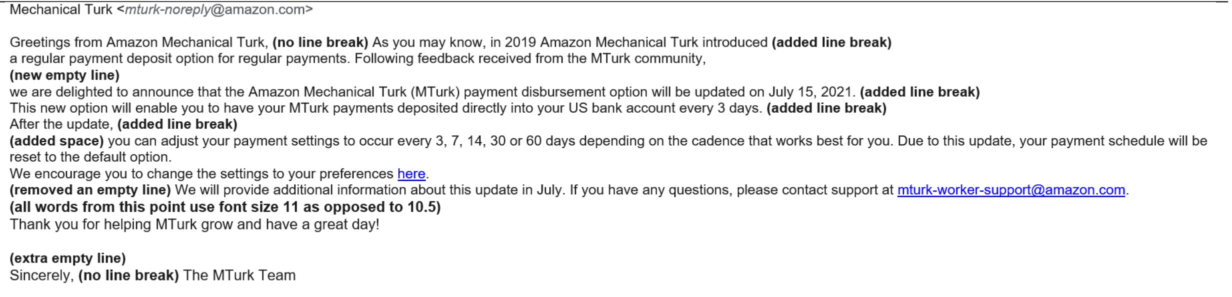
[29] Emma J. Williams and Danielle Polage. 2018. How persuasive is phishing email? The role of authentic design, influence and current events in email judgements. *Behaviour Information Technology* 38, 2 (Sept. 2018), 184–197. https://doi.org/10.1080/0144929X.2018.1519599

[30] Ryan T. Wright, Matthew L. Jensen, Jason Bennett Thatcher, Michael Dinger, and Kent Marett. 2014. Research Note: Influence Techniques in Phishing Attacks: An Examination of Vulnerability and Resistance. *Information Systems Research* 25, 2 (2014), 385–400. http://www.jstor.org/stable/24700179

Table 7: Appendix: Experiment Treatments

B=Baseline, T=Technical, C=Contextual, Lg=Language and Tone, Ly=Layout

The table below shows the treatments done for each factor on the baseline, to change a factor from value *high* to *low*. Note that as the experiment uses fractional factorial design, the specific changes shown for a factor in the table are never implemented alone, rather combinations of the treatments shown in the table have been implemented for the experiment. The parts that have been deleted from the baseline is denoted by striking through the deleted parts. The parts that have been added on top of the baseline is shown in bold text.

Factor	Email
B	Mechanical Turk <*mturk-noreply*@amazon.com> Greetings from Amazon Mechanical Turk, As you may know, in 2019 Amazon Mechanical Turk introduced a regular payment deposit option for regular payments. Following feedback received from the MTurk community, we are delighted to announce that the Amazon Mechanical Turk (MTurk) payment disbursement option will be updated on July 15, 2021. This update will enable you to have your MTurk payments deposited directly into your US bank account every 3 days. After the update, you can adjust your payment settings to occur every 3, 7, 14, 30 or 60 days depending on the cadence that works best for you. Due to this update, your payment schedule will be reset to the default option. We encourage you to change the settings to your preferences here. We will provide additional information about this update in July. If you have any questions, please contact support at mturk-worker-support@amazon.com. Thank you for helping MTurk grow and have a great day! Sincerely, The MTurk Team
T	~~Mechanical Turk <*mturk-noreply*@amazon.com>~~ **MTurk <*mtruk-noreply*@gmail.com>** Greetings from Amazon Mechanical Turk, As you may know, in 2019 Amazon Mechanical Turk introduced a regular payment deposit option for regular payments. Following feedback received from the MTurk community, we are delighted to announce that the Amazon Mechanical Turk (MTurk) payment disbursement option will be updated on July 15, 2021. This update will enable you to have your MTurk payments deposited directly into your US bank account every 3 days. After the update, you can adjust your payment settings to occur every 3, 7, 14, 30 or 60 days depending on the cadence that works best for you. Due to this update, your payment schedule will be reset to the default option. We encourage you to change the settings to your preferences here. We will provide additional information about this update in July. If you have any questions, please contact support at mturk-worker-support@amazon.com. Thank you for helping MTurk grow and have a great day! Sincerely, The MTurk Team
C	Mechanical Turk <*mturk-noreply*@amazon.com> Greetings from Amazon Mechanical Turk, ~~As you may know, in 2019 Amazon Mechanical Turk introduced a regular payment deposit option for regular payments. Following feedback received from the MTurk community,~~ We are delighted to announce that the Amazon Mechanical Turk (MTurk) payment disbursement option will be updated ~~on July 15, 2021~~ **soon**. This ~~update~~ **new option** will enable you to have your MTurk payments deposited directly into your bank account ~~every 3 days~~ **on a regular schedule**. ~~After the update,~~ You can adjust your payment settings to occur every 3, 7, 14, 30 or 60 days depending on the cadence that works best for you. Due to this update, your payment schedule will be set to the default option. We encourage you to change the settings to your preferences here. Thank you for helping MTurk grow and have a great day! Sincerely, The MTurk Team
Lg	Mechanical Turk <*mturk-noreply*@amazon.com> ~~Greetings from Amazon Mechanical Turk,~~ **Hi,** As you may know, in 2019 Amazon Mechanical Turk introduced a regular payment deposit option for regular payments. ~~Following feedback received from the MTurk community, we are delighted to announce that~~ **We hear you!** The new Amazon Mechanical Turk (MTurk) payment disbursement option will be updated on July 15, 2021**!** After this update you ~~can adjust your payment settings to occur every 3, 7, 14, 30 or 60 days depending on the cadence that works best for you.~~ **have the option to be paid every 3 days on top of the normal 7, 14, 30 or 60 day!** Due to this update, your payment schedule will be reset to the default option. You can Change your preferences here**!** We will provide additional information about this update in July. If you have any questions, send a mail to mturk-worker-support@amazon.com. Thank you for helping MTurk grow and have a great day!**!!** ~~Sincerely,~~ **Cheers,** The MTurk Team
Ly	Mechanical Turk <*mturk-noreply*@amazon.com> Greetings from Amazon Mechanical Turk, **(no line break)** As you may know, in 2019 Amazon Mechanical Turk introduced **(added line break)** a regular payment deposit option for regular payments. Following feedback received from the MTurk community, **(new empty line)** we are delighted to announce that the Amazon Mechanical Turk (MTurk) payment disbursement option will be updated on July 15, 2021. **(added line break)** This new option will enable you to have your MTurk payments deposited directly into your US bank account every 3 days. **(added line break)** After the update, **(added line break)** **(added space)** you can adjust your payment settings to occur every 3, 7, 14, 30 or 60 days depending on the cadence that works best for you. Due to this update, your payment schedule will be reset to the default option. We encourage you to change the settings to your preferences here. **(removed an empty line)** We will provide additional information about this update in July. If you have any questions, please contact support at mturk-worker-support@amazon.com. **(all words from this point use font size 11 as opposed to 10.5)** Thank you for helping MTurk grow and have a great day! **(extra empty line)** Sincerely, **(no line break)** The MTurk Team