IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 November 2024, accepted 5 December 2024, date of publication 11 December 2024,
date of current version 19 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3514603

STANDARDS

Quantifying Psychological Sophistication of Malicious Emails

THEODORE TANGIE LONGTCHI[1], ROSANA MONTAÑEZ RODRIGUEZ[2], KORA GWARTNEY[1], EKZHIN EAR[1], DAVID P. AZARI[3], CHRISTOPHER P. KELLEY[3], AND SHOUHUAI XU[1], (Senior Member, IEEE)

[1]Department of Computer Science, University of Colorado Colorado Springs, Colorado Springs, CO 80918, USA
[2]Department of Computer Science, University of Texas at San Antonio, Sans Antonio, TX 78249, USA
[3]Department of Behavioral Sciences and Leadership, United States Air Force Academy, Colorado Springs, CO 80840, USA

Corresponding author: Shouhuai Xu (sxu@uccs.edu)

This work was supported in part by NSF, a DoD UC2 Grant, under Grant #2115134; and in part by Colorado State Bill 18-086.

ABSTRACT Malicious emails (including Phishing, Spam, and Scam) are one significant class of cyber social engineering attacks. Despite numerous defenses to counter them, the problem remains largely open. The ineffectiveness of current defenses can be attributed to our superficial understanding of the psychological properties that make these attacks successful. This problem motivates us to investigate the psychological sophistication, or *sophistication* for short, of malicious emails. We propose an innovative framework that accommodates two important and complementary aspects of sophistication, dubbed *Psychological Techniques* (PTechs) and *Psychological Tactics* (PTacs). We propose metrics and grading rules for human experts to assess the sophistication of malicious emails via the lens of these PTechs and PTacs. To demonstrate the usefulness of the framework, we conduct a case study based on 1,036 malicious emails assessed by four independent graders. Our results show that malicious emails are psychologically sophisticated, while exhibiting both commonalities and different patterns in terms of their PTechs and PTacs. Results also show that previous studies might have focused on dealing with the less proliferated PTechs (e.g., Persuasion) and PTacs (e.g., `Reward`), rather than the most proliferated PTechs (e.g., Attention Grabbing and Impersonation) and PTacs (e.g., `Fit & Form` and `Familiarity`) that are identified in this study. We also found among others that social events are widely exploited by attackers in contextualizing their malicious emails. These findings could be leveraged to guide the design of effective defenses against malicious emails.

INDEX TERMS Cybersecurity metrics, cyber social engineering attacks, malicious emails, psychological factors, psychological tactics, psychological techniques, psychological sophistication.

I. INTRODUCTION

Cyber social engineering attacks have become an effective weapon for attackers to gain entry points into networks [1]. The consequences of these attacks have motivated many studies on designing countermeasures against them (e.g., [2], [3], [4], [5], [6], [7], [8], [9], [10]). Despite numerous efforts to stop these attacks, the attacks remain effective. Known defenses are often based on automated identification, typically via machine learning techniques (e.g., recognizing known Phishing email patterns). However, they do not consider psychological factors that can be leveraged to identify new approaches to defend against these attacks [11]. As suggested in [11], the initial step to effectively defending against these attacks is to systematically understand, characterize, and quantify the role of various psychological factors that contribute to their success. In addition, the industry has made many products to counter these attacks, such as FireEye Email Security, Mimecast, Cisco Cloud Email Security, Barracuda Sentinel, Microsoft Defender, and Norton LifeLock. Despite all these efforts, these attacks remain effective. For example, the 2022 Anti-Phishing

The associate editor coordinating the review of this manuscript and approving it for publication was Daniel Augusto Ribeiro Chaves.

Working Group (APWG) report [12] states that the number of Phishing attacks reported to APWG has quadrupled since early 2020; a more recent report [13] states that 2023 has been the worst year on record so far, with more than 5 million Phishing attacks.

The preceding discussion inspires us to investigate the notion of *psychological sophistication*, or *sophistication* for short, of malicious emails to pave a way towards designing effective defenses. Our focus is on investigating email sophistication with respect to the psychological elements in malicious emails that attackers employ to lure victims. That is, our focus is on individual email and its content, which include the sender, the recipient, the day of the year, the subject of the email, and the inner presentation of the email such as the salutation, the body, the signature, and the footer. To our knowledge, these psychological elements have not been systematically studied in the literature. This is true despite the studies that considered the impact of the following non-psychological factors on the success of Phishing attacks: the time of the day, the day of the week, or the month of the year that an email is sent [14], [15], [16].

Our Contributions: This paper makes three contributions. First, we initiate a systematic investigation on the psychological sophistication of malicious emails. Specifically, we ask three questions: (i) How can we quantify the psychological sophistication of malicious emails in the real world? (ii) What kinds of psychological sophistication patterns are exhibited by different kinds (or types) of malicious emails in the real world? (iii) How does the sophistication of malicious email evolve over time?

Second, we propose an innovative and systematic framework to quantify the psychological sophistication of malicious emails. The innovation of the framework is to deconstruct low-level and high-level psychological features that influence an email recipient to engage with the malicious content. The framework deconstructs and compares the content of malicious emails through two lenses:

- *Psychological Techniques (PTechs)*: At the low level, we propose identifying the number of psychologically relevant textual and imagery elements in an email message, dubbed *Psychological Techniques* (PTechs) to provide a detailed accounting of the elements employed in implementing an attack.
- *Psychological Tactics (PTacs)*: At the high level, we propose assessing the attacker's overall deliberate thoughtfulness (reflecting attacker's effort) in framing malicious content to influence an email recipient, dubbed *Psychological Tactics* (PTacs), to measure an attacker's effort in exploiting human fallibility.

We propose considering both PTechs and PTacs because they offer complimentary views of an attacker's effort as reflected by the observed elements in malicious emails. Together, they enable a rich exploration of qualitative and quantitative insights into how an attacker frames and implements its attack. As a side-product, the framework could be adapted / adopted to quantify the sophistication of other types of cyber social engineering attacks, such as malicious text messages.

Third, we demonstrate the usefulness of the framework by applying it to quantify the sophistication of 1,036 malicious emails. This empirical study allows us to draw useful insights, including: (i) previous studies might have focused on dealing with the less proliferated PTechs (e.g., Persuasion) and PTacs (e.g., `Reward`), rather than the most proliferated PTechs (e.g., Attention Grabbing and Impersonation) and PTacs (e.g., `Fit & Form` and `Familiarity`) that are identified in the present study; (ii) Phishing emails are psychologically more sophisticated than Spam and Scam emails; (iii) PTechs are independently employed by attackers, suggesting no coordination between attackers; (iv) malicious emails that are sophisticated in PTechs are also sophisticated in PTacs.

As a side-product, we present the algorithm for computing the degree of agreement between graders on a given set of objects (i.e., emails in this context) with respect to multiple attributes (i.e., PTechs and PTacs in this context). The idea behind the algorithm was presented in the literature [17], [18], [19], but we are not aware of any other algorithmic description in the literature. Thus, the algorithm may be of independent value to the computer science community.

Ethical Issue: In consultation with the University of Colorado Colorado Springs Internal Review Board (IRB), this study does not need IRB approval because no subjects are part of the study and the emails are provided by third parties.

Related Work: Longtchi et al. [11] systematize the psychological aspects relevant to cyber social engineering attacks. The present study is inspired by [11], and proposes the first framework for quantifying psychological sophistication of malicious emails. The framework can be incorporated into the broader framework presented in [11] towards ultimately tackling cyber social engineering attacks.

Studies have investigated the use of psychological content in Phishing emails (e.g., [7], [20], [21], [22], [5], [23]). Allodi et al. [20] study Phishing and Spear Phishing and defenses, showing that SCARCITY is employed in Phishing emails, but defenses do not account for human-related characteristics. Goel et al. [7] investigate humans' susceptibility to *deception*. Flores et al. [21] study correlation between personal psychology and demographics in terms of resistance to Phishing, finding that neither age nor gender significantly correlates to Phishing resilience but computer experience does. The other studies examine the psychological content of social engineering messages through the lens of Cialdini's Persuasion Principles and other principles. Heijden and Allodi [22] study the identification of persuasion elements in Phishing emails. Nelms et al. [5] investigate the psychological tactics to encourage users to download malicious applications, by considering Whaley's Theory of Deception in addition to Cialdini's Persuasion Principles. Ferreira and Lenzini [23] investigate the psychological content in Phishing messages based on low-level psychological elements by leveraging Cialdini's Persuasion Principles, Stajano's Scam

Principles, and Gragg's psychological triggers. In particular, we consider both low-level and high-level psychological features. De Bona and Paci [24] show, among other things, that urgency is more effective than authority in making employees susceptible to Phishing attacks. By contrast, we systematically study the impact of PTechs and PTacs on the sophistication of malicious emails.

Studies have investigated the use of psychological content in Spam emails. Gallo et al. [25] design a system that can detect persuasive elements in Phishing email, such as consistency, Urgency, SCARCITY, LIKING, SOCIAL PROOF, AUTHORITY, and RECIPROCITY. Gallo et al. [26] investigate Spam emails collected in a 2-year span and show how Phishing emails exploit cognitive vulnerabilities, such as the use of deceiving words (e.g., *account, suspended* and *verify*) or scamming words (e.g., *donate*, *please*, and *warning*). Vance et al. [27] study the effect of habituation and generalization with respect to warnings and notification signs in an online setting, where habituation indicates people's response to repeated stimulation decreases over time and generalization means habituation to a stimulus is carried over to other stimuli that are similar in appearance. However, these studies do not consider the quantification of psychological sophistication of malicious emails.

To determine how personality traits influence susceptibility to social engineering attack, Uebelacker and Quiel [28] propose a Social Engineering Personality Framework (SEPF) to describe the relationships between the personality traits of the Big Five Model and the Cialdini's principles of influence. The framework aims to support and guide security researchers and practitioners in developing detection, mitigation, and prevention strategies while dealing with human factors in social engineering attacks. While this framework and ours are both geared towards providing insights in social engineering attacks for a better defense system, our paper conducts a case study to test the usefulness of the framework, which lead to some important insights. Relevant prior studies (e.g., [29], [30], [22], [23], [31], [32], [33]) that investigate how PTechs are exploited by malicious emails only consider one or very few PTechs. For example, Wang et al. [33] study the impact of visceral triggers on Phishing susceptibility, but only considering few visceral triggers (e.g. fear, the stressing of Urgency to respond, or the implicit use of Impersonation). By contrast, we consider the evolution of 7 PTacs and 8 PTechs from 2006 to 2022 with intervals of 5 years (i.e., 2006, 2011, 2036, and 2021 plus 2022) using a large dataset of 1,036 malicious emails. Longtchi and Xu [34], [35] study the evolution of psychological factors exploited by malicious emails and the evolution of 9 PTechs and 7 PTacs over the past 21 years (2004-2024). These studies can be extended to characterize the evolution of psychological sophistication of malicious emails by leveraging the metrics introduced in the present study. When compared with all the studies mentioned above, we propose a systematic framework for quantifying the psychological sophistication of malicious emails.

Paper Outline: Section II describes the psychological concepts used in the paper. Section III presents the framework for quantifying sophistication of malicious emails. Section IV reports a case study on applying the framework to a malicious email dataset. Section V discuses limitations of the study. Section VI concludes the paper with research directions.

II. THE CONCEPTS OF PTechs AND PTacs

A. BASIC IDEA

We propose using *low-level* and *high-level* psychological features to characterize sophistication of malicious emails. Low-level psychological features are visible representations of salient textual or imagery elements in a (malicious) email to increase the likelihood of recipient compliance. These elements could resemble a familiar picture, logo, or keyword that instills a sense of confidence in the recipient. High-level features, by contrast, reflect the overall email framing (i.e., presentation). For example, sending a message from a (purported) source of authority (e.g., a supervisor) while amplifying pressure to act under time pressure (e.g., "We have a problem - call me immediately"). The low-level features (i.e., elements) and high-level features (i.e., framing) can be seen as the respective counterparts of the attack *techniques* and *tactics* in MITRE's ATT&CK framework [36]. Thus, we refer to the low-level psychological features as *Psychological Techniques* (PTechs) and high-level psychological features as *Psychological Tactics* (PTacs). We use the concepts of PTechs and PTacs to define metrics and quantify sophistication of malicious emails.

B. PTechs

A PTech is a concrete (i.e., quantifiable) cue such as a textual or an imagery element that encourages individuals to comply with a social engineering attack. PTechs that have been identified in the literature include [11], [37]:

1) Urgency: The use of textual elements (e.g., "acting now") to trigger a quick response from the recipient [8], [38]. That is, putting a time constraint on recipients to force them to act fast without thoughtfulness.
2) Visual Deception: The use of visual elements (e.g., logos) or "similar" characters in URL (e.g., replacing 'vv' with 'w', or 'm' with 'rn') to project trust [39], [40]. The attacker's malice is hidden in plain sight.
3) Incentive & Motivator: The use of textual or graphic elements to indicate a high discount or a freebie such as "gift cards" (incentive) or "help others", to incentivize or motivate a recipient to take action [37], [41].
4) Persuasion: The use of textual elements related to Cialdini's principles of persuasion (e.g., "C-Suite titles," "last chance," or "expert opinion") to encourage a recipient to take action [23], [30]. The six principles are: AUTHORITY, which describes power or dominance over someone; RECIPROCATION (or RECIPROCITY), which describes the tendency to pay back a favor; LIKING (or SIMILARITY), which

describes one's tendency to react positively to those people they have a relationship; SCARCITY, which uses the lack of goods or services to lure victims; SOCIAL PROOF, which describes one's tendency to imitate others; CONSISTENCY (or COMMITMENT), which describes the extent one is dedicated to a person or something.

5) **Quid-Pro-Quo**: The use of textual elements (e.g., "Pay an upfront fee") to ask a recipient for a favor in exchange for a bigger reward [42]. Note that **Quid-Pro-Quo** is different from RECIPROCITY because the former is about agreeing to pay back before a deed (i.e., similar to palm-greasing or bribery) but the latter is about paying back after a deed (i.e., similar to freewill).
6) **Foot-in-the-Door**: The use of textual elements (e.g., "from our last email ...") to obtain compliance from a recipient via gradually increasing demands [43]. Note that it is different from politeness because it eventually gains full access by making an individual to gradually accept modest requests.
7) **Trusted Relationship**: The exploitation of an established third-party relationship of trust with the recipient by using textual elements, such as "John told me about you" to convince a recipient to take action [20].
8) **Impersonation**: The use of a false persona to gain the trust of a recipient by using elements, such as "I'm billionaire Warren Buffet" [20], [23]. Note that **Impersonation** is different from **Pretexting** because the former takes a fake persona to hide one's real identity or to gain trust but the latter presents a fake narrative/story to gain trust.
9) **Contextualization**: Referencing current event by using textual elements, such as "the Pandemic" or "War in Ukraine" [3], [7], [40]. For example, consider an email that contains the following sentence: *"we are collecting donations for the victims of the Ukraine war who migrated to the USA, in order to help them celebrate Thanksgiving."* In this email, two contexts are exploited by the attacker: the war in Ukraine and Thanksgiving. In general, the events that are exploited can be an activity that is only known to members of a closed community (e.g., work retreat).
10) **Pretexting**: Providing a motive to establish contact with a recipient by using textual elements, such as "I am recruiter for XYZ company" [7], [44].
11) **Personalization**: Addressing a recipient using detailed personal information in textual elements, such as "Dear John" or "Your credit card ending in ..." [45], [46].
12) **Attention Grabbing**: The use of graphical/auditory elements to draw attention to textual elements, such as highlighted text, brightly colored buttons, or extra large fonts [5], [21]. Note that unlike **Visual Deception** where attackers use visuals to instill trust, attackers use visuals in **Attention Grabbing** to draw a recipient's attention to what the attacker wants the recipient to see or do.
13) **Affection Trust**: Developing an effective relationship to extort a recipient by using textual elements, such as "My child is sick and I have no money to pay for the treatment" [47].
14) **Decoy Effect**: Making one to believe that something is a good deal (e.g., presenting a user with a lower than the market price for some goods but actually offering a fake one or never delivering when a victim pays upfront) [48], [49].
15) **Priming**: Influencing one's decision through gradual manipulation (e.g., sending them information about cryptocurrency as the next big thing before sending them a fake link to purchase cryptocurrency) [50].
16) **Loss Aversion**: Providing something for free, but later charging enormously when a victim becomes attached to the free item (e.g., providing live soccer links, then charging them when they become attached to the free live soccer links) [51].

C. PTacs

A PTac is a measure of the overall coherence and quality of the message based on established ideas of framing and relevance [7] to influence decision-making [40]. Each PTac aims to measure an attacker's effort at crafting and framing an email effectively to prompt a recipient's action [52]. PTacs can be extended to accommodate other psychological factors and framing approaches that may be proposed in future studies. PTacs that have been implicitly, but not explicitly, proposed in the literature include:

1) `Familiarity`: This refers to how an attacker engenders a positive (and therefore trusting) association with a recipient. Emails of high `familiarity` may impersonate specific people (e.g., co-workers, bosses, family members, close friends) [44], [47].
2) `Immediacy`: This refers to the amplification of a time constraint as a mechanism to shortcut a recipient's skepticism or scrutiny for a desired action, for example, by suggesting that promptness, swiftness, or a quick reaction is required [5], [40]. Unlike the **Urgency** PTech that counts the number of instances of elements that can trigger a quick response are employed in an email, while `Immediacy` assesses and scores the attacker's overall effort in employing elements that trigger a quick response in an email. For example, if an email uses "now", "now", "now" three times, it will be 3 counts of the **Urgency** PTech. However, the `Immediacy` PTac evaluates the overall effect of those 3 instances of urgency in the email. Therefore, a value of 3 for the **Urgency** PTech in an email is not necessarily a value of 3 for `Immediacy`. This is because the latter considers the other elements of the email as well (e.g., where are the 3 "now" placed in the email? Are they placed together or separately in places

of the email such the subject line, the body, or closing line?).

3) `Reward`: This refers to a clear exchange of something (physical or social) valuable for a recipient. Rewards are often presented as tangible goods (e.g., money) in exchange for action but can be an offer to improve social standing (e.g., authority, prestige) [7], [11].
4) `Threat of Loss`: This refers to an appeal to a recipient's desire to maintain their current status, prevent a loss (e.g., opportunity) or injury (e.g., damage, pain), or avoid the risk of having something stolen. Loss has been hypothesized to be more impactful than potential gain (e.g., reward) [7], [23], [42].
5) `Threat to Identity`: This refers to the efforts by an attacker to manipulate a recipient's desire to maintain a positive, socially valuable reputation (e.g., "Pay your dues or face the consequences") [42], [47].
6) `Claim to Legitimate Authority`: This refers to emphasizing a source of legitimate power to obscure or deter increased scrutiny. The attacker may assume a position of technical expertise, a valuable institutional role, or a traditionally respected office [23], [42]. Note that it is different from the Principle of Authority because a legitimate authority in the former case does not have to be a human but the authority in the latter case refers to human experts by definition.
7) `Fit & Form`: This refers to how a message mirrors the expected composition style of an authentic message. An attacker often exploits commonly expected written or visual display format to resonate with the email's apparent sender and purpose [7], [53]. Note that unlike `Familiarity` which is about being familiar with the content of an email, `Fit & Form` is the general expectation on how the email should look and feel when coming from the purported sender.

III. FRAMEWORK

The framework is centered around using PTechs and PTacs to measure the sophistication of malicious emails. Intuitively, a malicious email is an email whose objective is hidden from the recipient but beneficial to the sender (i.e., attacker). In other words, a malicious email attempts to coerce a recipient to do something in compliance with the request described in the email, which the recipient would not do if the recipient knew the sender's objective. The rationale is that each psychological feature represents a different aspect of the attacker's effort.

Intuitively, PTechs can be seen as *quantifiers* of a malicious email content, reflecting the presence of *elements* that evidence the attackers' effort; whereas, PTacs can be seen as *qualifiers* of malicious email content, reflecting the attackers' overall effort. To our knowledge, this is the first work that provides a systematic decomposition of email contents coupled with attacker effort to quantify the sophistication of malicious emails. Moreover, the framework can be extended to accommodate new factors that may be identified by future advancement in psychology research.

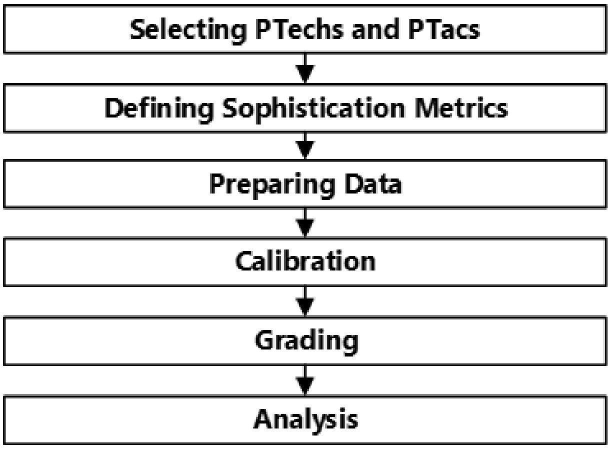


FIGURE 1. Overview of the framework, where the calibration process is iterative.

The framework consists of six components as shown in Figure 1: (i) selecting PTechs and PTacs for assessment; (ii) defining metrics to quantify sophistication of malicious emails; (iii) preparing a dataset of malicious emails for expert graders to assess; (iv) calibrating the grading process, including designing grading rules and training; (v) grading emails; and (vi) analyzing outcome of the grading process. These steps may look unfamiliar to computer scientists, who are often given datasets to analyze.

A. SELECTING PTechs AND PTacs

We propose selecting the PTechs that (i) are known to be used in malicious emails based on research evidence and (ii) require a one-time interaction to be effective. This selection criterion is flexible enough to accommodate future understanding and knowledge (e.g., when new PTechs are discovered in the future). Similarly, we propose selecting PTacs that (i) are known to be used in malicious emails based on research evidence, (ii) are independent of one another, and (iii) reflect the holistic effort of an attacker. Suppose, according to the respective selection criteria, some number ℓ of PTechs are selected, denoted by $\{\text{PTech}_1, \ldots, \text{PTech}_\ell\}$, and some number m of PTacs are selected, denoted by $\{\text{PTac}_1, \ldots, \text{PTac}_m\}$. This allows us to define sophistication metrics as follows.

B. DEFINING SOPHISTICATION METRICS

We propose expressing sophistication of malicious emails through PTech-based and PTac-based analysis and quantifying it as a two-dimensional vector. We consider both the "ideal world" where there are no outliers in terms of subjective measurements and the "real world" where there are outliers. This is important and inevitable because we need human experts to "grade" malicious emails with respect to PTechs and PTacs. This process is analogous to obtaining the ground-truth labels in machine learning, namely that we need ground-truth labels to train models. Moreover, outliers could still be relevant even after having some calibration process that aims to train human graders in agreeing on how to grade. The calibration process is one component of the framework and will be described later.

1) METRICS FOR MEASURING PTechs

To clarify our way of thinking, let us start with the ideal world. Consider a given malicious email and a set of ℓ PTechs denoted by $\{\text{PTech}_1, \ldots, \text{PTech}_\ell\}$. For PTech_i where $1 \leq i \leq \ell$, we propose counting the number of elements with respect to PTech_i, leading to an integer grade (or score) s'_i. Assuming the s'_i for every i is defined in the same range $[0, \gamma] = \{0, 1, \ldots, \gamma\}$. Then, the sophistication of the malicious email through the lens of the ℓ PTechs can be defined as,

$$s' = \frac{1}{\ell} \sum_{i=1}^{\ell} s'_i.$$

In the real world, the ground-truth s'_i is difficult to obtain. Thus, we propose approximating it by using a number of n graders (or evaluators) to count the elements concerning PTech_i while assuring that the graders can count the elements as consistently as possible. For a given malicious email, let $s_{i,j}$ denote the count of elements in the email by grader j with respect to PTech_i, where $1 \leq j \leq n$ and $1 \leq i \leq \ell$. Then, the sophistication of the email concerning PTech_i can be defined as,

$$S_i = \frac{1}{n} \sum_{j=1}^{n} s_{i,j}. \tag{1}$$

Given S_i for $1 \leq i \leq \ell$, we propose defining the sophistication of the given malicious email with respect to the ℓ PTechs, denoted by S_{PTech}, as:

$$S_{\text{PTech}} = \frac{1}{\ell} \sum_{i=1}^{\ell} S_i. \tag{2}$$

Note that Eq.(1) operates by assuming every grade should be considered. In the practice, some grade(s) by some grader(s) may be considered outliers and thus eliminated. This means that Eq.(1), should be amended to accommodate the elimination of outliers, while keeping Eq.(2) intact. Specifically, when coping with Eq.(1), which computes the average grade S_i of PTech_i by the n graders, we may encounter a subset of graders, denoted by $J' \subset \{1, \ldots, n\}$, being outliers according to some well-established criteria (e.g., the specific one that will be used in our case study, but there could be others), meaning that their grades, namely $s_{i,j}$ for $j \in J'$, should be excluded when computing the average grade S_i. (Note that $J' = \emptyset$ corresponds to the absence of outliers). As a result, Eq.(1) becomes for $1 \leq i \leq \ell$:

$$S_i = \frac{1}{n - |J'|} \left(\sum_{j=1}^{n} s_{i,j} - \sum_{j \in J'} s_{i,j} \right). \tag{3}$$

2) METRICS FOR MEASURING PTacs

Similarly, we start with the ideal world. Consider a malicious email and a set of m PTacs denoted by $\{\text{PTac}_1, \ldots, \text{PTac}_m\}$. Since the ground-truth sophistication is difficult to obtain, we propose assessing PTac_i using a scale $[0, \beta] = \{0, 1, \ldots, \beta\}$ for some integer β, also by n independent graders, where $p_{i,j}$ denotes the assessment of grader j with respect to PTac_i for a given malicious email, $1 \leq j \leq n$, and $1 \leq i \leq m$. The sophistication of the given malicious email with respect to PTac_i can be defined as,

$$P_i = \frac{1}{n} \sum_{j=1}^{n} p_{i,j}. \tag{4}$$

The overall PTac-based sophistication of an email can be defined as:

$$S_{\text{PTac}} = \frac{1}{m} \sum_{i=1}^{m} P_i. \tag{5}$$

Note that Eq.(4) assumes there are no outlier grades, which need to be eliminated. Denote by $J'' \subset \{1, \ldots, n\}$ the set of outliers, meaning that their grades, namely $p_{i,j}$ for $j \in J''$, should be excluded when computing the average grade P_i, where $J'' = \emptyset$ corresponds to the absence of outliers. As a result, Eq.(4) now becomes for $1 \leq i \leq m$:

$$P_i = \frac{1}{n - |J''|} \left(\sum_{j=1}^{n} p_{i,j} - \sum_{j \in J''} p_{i,j} \right), \tag{6}$$

but Eq.(5) remains unchanged.

3) METRICS FOR QUANTIFYING SOPHISTICATION OF MALICIOUS EMAILS

By treating $J' = \emptyset$ as a special case of $J \neq \emptyset$ and treating $J'' = \emptyset$ as a special case of $J'' \neq \emptyset$, we obtain:

Definition 1 (Sophistication of Malicious Email): The sophistication of a malicious email is measured as a two-dimensional vector (S_{PTech}, S_{PTac}), where S_{PTech} is defined in Eq.(2) with S_i in Eq.(2) being defined in Eq.(3), and S_{PTac} is defined in Eq.(5) and with P_i in Eq.(5) being defined in Eq. (6).

C. PREPARING DATA

Several issues must be addressed when preparing data, including collection and preprocessing. First, to ensure dataset quality, we must assure the emails are malicious as the purpose is to quantify their psychological sophistication and show the importance and relevance of considering psychological sophistication in future studies (e.g., different defensive mechanisms may be used to cope with malicious emails with different degrees of psychological sophistication).

Second, we must ensure that the data preparation process does not cause damage to the research environment. This is important because malicious emails may contain links to executable code or malicious websites that can compromise the experimental environment if accidentally clicked, thus possibly affecting other computers in the network. Therefore, using a virtual environment to isolate emails from the Internet is imperative when preprocessing malicious emails.

Third, malicious emails in a given dataset may contain broken links or missing images, which are needed for assessing their sophistication because these contents would be presented when a recipient views an email in the real world. This means that we must reconstruct an email by adding the missing links or images. For example, if a malicious email is missing the Amazon.com logo, we can reconstruct the email by adding the mission logo. If we cannot reconstruct an email, the email should not be used in this study. Moreover, the reconstruction process should be sound. For example, an email dated with year 2006 with a missing PayPal logo must be replaced only with the 2006 PayPal logo.

Fourth, given a set of malicious emails, we must ensure that each email content is rendered similarly, if not exactly the same, on different machines and platforms from a visual point of view. This is important because an email will be assessed by multiple graders. This is not trivial to guarantee because graders may use different software platforms, email readers, or web browsers, which might render emails differently on each computer. We ensure this by presenting screenshots of emails to the graders.

D. CALIBRATION

The calibration process aims at mitigating human (including expert) subjectivity in grading sophistication of malicious emails. In a sense, the calibration process is reminiscent of the model training step in machine learning. The calibration process has two sub-processes: *Designing Grading Rules* and *Training*, which are highlighted in Figure 2, including the iteration that may be incurred. Calibration is important because as mentioned above, we can only approximate the unknown ground-truth metrics by having domain experts grade emails with respect to PTechs and PTacs according to our *grading rules* (in a fashion similar to [54], [55], [56]). When a grader manually counts the number of psychological elements exhibited in an email with respect to a PTech and assesses the overall coherence and quality of the email with respect to a PTac, bias and/or subjectivity is inevitable because the interpretation of "psychological elements" relies on one's domain expertise. This explains why we need the *Designing Grading Rules* sub-process. Even given grading rules, we still need to train graders to eliminate as much bias or subjectivity as possible.

1) THE DESIGNING GRADING RULES SUB-PROCESS

The purpose of grading rules is to guide graders in scoring the psychological sophistication, more specifically psychological elements, of emails with respect to PTechs and PTacs. We propose designing grading rules in an iterative fashion as follows.

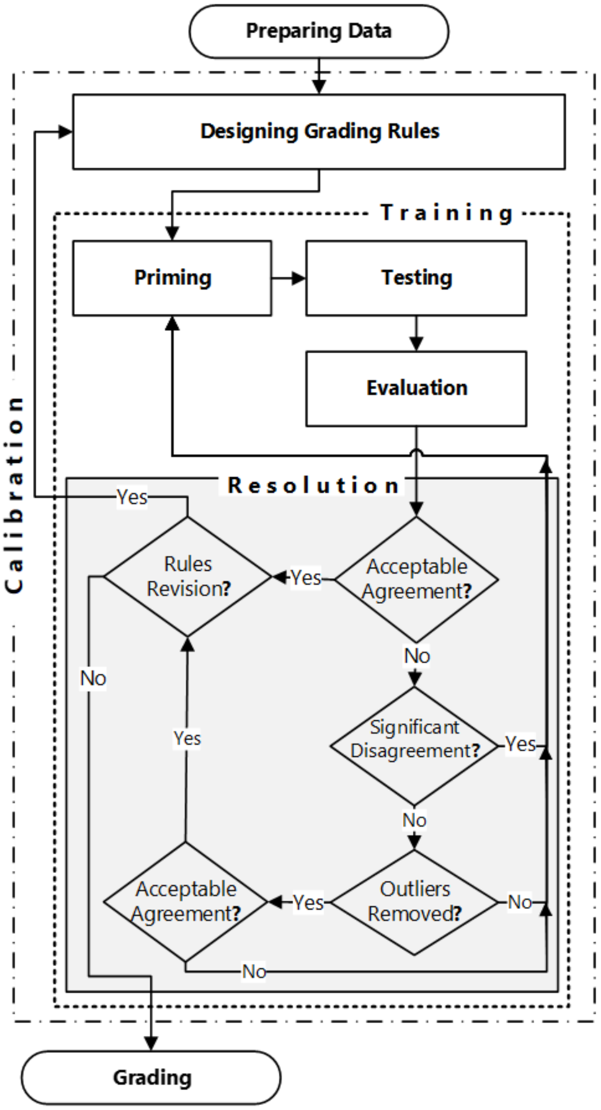


FIGURE 2. The Calibration Process includes two sub-processes: *Designing Grading Rules* and *Training*, where the latter has four steps—priming, testing, evaluation, and resolution.

1) Initial grading rules are designed by some experts, such as: (i) instructions on recognizing a PTech, how to count its use, and what the metric scale is; (ii) instructions on scoring the effectiveness of a PTac in an email, and what is the metric scale. In our case study, we will present specific initial grading rules with respect to each PTech and PTac.
2) The initial rules are used in the *Training* sub-process, which (as described below) helps decide whether to revise the current grading rules (e.g., adding new rules or deleting existing rules). If no revision is necessary, the current rules will be used in the subsequent Grading process; otherwise, the current rules will be revised and

the revised rules will be used in the subsequent iteration of the Training sub-process (i.e., effectively returning to the previous step).

3) Rule for Removing / Keeping Outlier Grades. When there is a statistically significant variation of agreement among the grades, we may need to eliminate the outlier grades that are outside of a predetermined acceptable threshold. For this purpose, we need a rule to specify: (i) what makes a grade assigned to an email with respect to a specific PTech or PTac an outlier; (ii) how to decide whether an outlier will be definitely eliminated or not, which is relevant especially when the number n of graders is small.

Since grading rules may be complicated, a *grading aid* may be warranted. Intuitively, a grading aid would display the grading rules to the graders in their course of grading so as to reduce their cognitive load on conducting the task. This is important because a significant number of PTechs and PTacs are involved, and the relationships between these PTechs and PTacs can be delicate. For example, a grading aid may provide a reference table of key psychological elements corresponding to a PTech or PTac and may further provide examples of each PTech or PTac using example emails.

2) THE TRAINING SUB-PROCESS

Given the initial or revised grading rules, a group of graders are recruited. The graders learn how to apply these rules to grade emails in the *Training* sub-process, which has the following four steps.

1) **Priming**: At this step, graders learn how to grade emails using the initial or current grading rules, ideally supported by a grading aid as mentioned above. Graders can ask each other questions (e.g., what would count as a psychological "element" with respect to a PTech?), resolve disagreements between them, and collectively build a shared understanding of the assessment method. Graders may collectively grade sample emails and discuss them. Then, the sub-process moves to the *Testing* step below.
2) **Testing**: At this step, each grader assesses a set of emails independently. The grading environment used in this step should be the same as the one that will be used in the grading process.
3) **Evaluation**: Grades (i.e., scores assigned by graders) resulting from the *Testing* step are evaluated for their degree of agreement, which can be measured with some well-established metrics. We advocate using Krippendorff's Alpha (Kalpha) [17], denoted by $\alpha \in [-1, 1]$, because it is not only applicable to categorical, ordinal, and interval data, but also robust in the presence of missing data (which can occur when outliers are removed) [19], [57], [58]. One advantage of using Kalpha over the *standard deviation*, which may be tempting, is that it is not clear how standard deviation can be used to measure the agreement among graders on many emails with respect to many PTechs and PTacs. Moreover, it is not known how standard deviation can account for "agreement by chance" among the graders, whereas Kalpha addresses this issue leveraging the ratio of the observed disagreement to the expected disagreement [19]. The meaning of the Kalpha (α) is interpreted as follows [17], [19]: $\alpha = -1$ means absolute disagreement; $-1 \leq \alpha < 0$ indicates a degree of disagreement; $0 < \alpha < 0.6$ means an unreliable degree of agreement and a resolution among the graders is required; $0.6 \leq \alpha < 0.8$ means an acceptable degree of agreement but there may still be a degree of disagreement which may need to be resolved; $0.8 \leq \alpha < 1$ means a highly reliable degree of agreement (i.e., agreement beyond chance and a solid conclusion can be made from the grades); $\alpha = 1$ means a perfect agreement. These parameter regimes will guide the use of α in the *Resolution* step below.
4) **Resolution**: At this step, a resolution is to be made based on the degree of agreement α obtained in the preceding step.
Corresponding to Figure 2, we consider three cases. (i) If $0 < \alpha < 0.6$, meaning there is a significant disagreement among the grades, the graders return to the *Priming* step of the Training sub-process. (ii) If $0.6 \leq \alpha < 0.8$, meaning there is a potentially acceptable degree of agreement but also a degree of disagreement, the graders may try to resolve the disagreement. One method is to eliminate outlier grades and see if the resulting α significantly increases. If eliminating outliers does not increase α significantly, the graders return to the *Priming* step. If eliminating outliers does increase α significantly, the graders need to decide whether to revise the grading rules or not. If not, they proceed to the *Grading* process; otherwise, they return to the *Designing Grading Rules* sub-process. (iii) If $\alpha \geq 0.8$, the grades are in highly agreement with each other and proceed to *Grading*.

E. GRADING

Having accomplished the calibration process, the graders are ready to grade emails (which are not used in the calibration process). In this process, the graders *independently* assess each email by assigning PTech grades and PTac grades according to the email, while leveraging the grading aid mentioned above (if present). The resulting grades will be used for analysis in the next step.

Though sounding simple, the grading process encounters some issues. First, we should not assume the graders can memorize the grading rules, which should be presented to the graders in the entire grading process. This can be done, for example, by making the grading process based on online survey, while presenting the grading rules or grading aid to the graders in an easy-to-reference fashion.

Second, we must deal with the issue of graders' fatigue. This is necessary because it is not feasible for each grader to grade a large number of emails with high quality within a consecutive period of time. This can be achieved, for example when designing the online survey mentioned above, by breaking down the grading into smaller sessions each of which only requests a grader to grade a small number of emails. Third, we must give the graders flexibility in terms of the time frame to finish grading because graders may have different working styles or schedules. Fourth, we must mitigate the bias that may be introduced in the grading process. For example, even if each grading session has a small number of emails, it is possible that the quality of grading by a grader may decrease somewhat as the grading proceeds (e.g., fatigue may play a role even within a single session of 100 emails). To mitigate this factor, we can shuffle the emails in terms of the order they are presented to the graders.

F. ANALYSIS

This process is to analyze the resulting grades from the *Grading* process. To make the analysis process more effective, it would be better to design a set of Research Questions (RQs). Examples of Research Questions (RQs) include the following (while noting that other RQs are possible):

- RQ1: Which PTechs and PTacs are widely employed?
- RQ2: What emails are more sophisticated?
- RQ3: How are PTechs and PTacs correlated?
- RQ4: Why emails are inconsistently graded?
- RQ5: Does sophistication evolve with time?
- RQ6: How do attackers exploit social events?

IV. CASE STUDY

Now we present a case study to demonstrate the usefulness of the framework, while using an end-to-end example to show the processes.

A. SELECTING PTechs AND PTacs

Under the guidance of the PTech and PTac selection criteria described in framework (Section III-A), we select the following 8 (out of the 16) PTechs described in Section II: **Urgency**, **Incentives & Motivators**, **Attention Grabbing**, **Personalization**, **Contextualization**, **Persuasion**, **Impersonation**, and **Visual Deception**. In the terminology of the framework, we have $\ell = 8$. Moreover, we select all the 7 PTacs described in Section II: `Familiarity`, `Immediacy`, `Reward`, `Threat of Loss`, `Threat to Identity`, `Claim to Legitimate Authority`, and `Fit & Form`. In the terminology of the framework, we have $m = 7$.

B. INSTANTIATING SOPHISTICATION METRICS

In order to use the sophistication metric given by Definition 1, which is made general enough to accommodate any metric scale, we need to instantiate the metric scales for the PTechs and PTacs, respectively. First, for each PTech, we propose using the scale $[0, 7] = \{0, 1 \ldots, 7\}$ or $\gamma = 7$, for two reasons. (i) When developing the grading aid, we notice that most elements of the PTechs do not go above 7 (with the only exception of **Attention Grabbing**). (ii) By limiting the maximum count of PTechs, we mitigate the damage that can be caused when one PTech dominates the PTech grade. Second, for each PTac, we propose using the Rating scale [0, 5] (i.e., by setting $\beta = 5$) which is commonly used in psychological studies [59].

C. PREPARING DATA

We focus on three types of malicious emails: (i) Phishing emails, which require a one-time interaction for victimization and include a link or an attachment; (ii) Scam emails, which require multiple interactions for victimization via phone call or email exchange, or request personal information; and (iii) Spam emails, which may not obscure information but often sells a product or service that can be misleading.

To obtain data (i.e., malicious emails), we collected emails from the APWG as follows. (i) We became an APWG member so that we can have access to the APWG database of malicious emails. (ii) We created an account, whereby we accessed the emails that were submitted to APWG by its members (e.g., financial institutions, retailers, Internet Service Providers or ISPs, solution providers, law enforcement, government agencies, university researchers, and Non-Government Organizations or NGOs). (iii) We selected emails using the APWG Reported Phishing module API. To increase the chance of collecting truly malicious emails, we selected emails submitted by US-CERT (Computer Emergency Response Team) as it is a reputable source. (iv) We reassembled the headers and the bodies of the select emails into `.eml` files using a Python code we wrote. This is because email headers and bodies are stored separately in the APWG database, but they can be linked together via email IDs. (v) We used an email client (i.e., Thunderbird) to read the emails and take screenshots.

We prepare data according to the framework as follows. First, to assure that the emails are truly malicious, we use reputable data sources. In total, we collect and use 1,177 emails from two sources:

- 1,143 emails from the Anti Phishing Working Group (APWG) repository [60], which collects malicious emails from various sources. These 1,143 emails are selected because they are submitted to APWG by US-CERT (Computer Emergency Response Team), which is a reputable source accepting reports of malicious emails and security incidences from government agencies as required by U.S. federal law [61]. Among the 1,143 emails, 107 will be used in the *Calibration* process and 1,036 will be used in the *Grading* process. The 107 emails used for calibration is roughly 9.4% of the 1,143 emails in volume for this study, while noting that the 1,143 – 107 = 1,036 emails will be used for actual analysis. We selected 9.4% of the emails for

calibration by loosely applying the 10% rule. In data analysis, the 10% rule states that if a sample size is less than 10% of a population, then the trials in the sample can be treated as independent [62], [63]. It is worth noting that we applied this rule just to statistically gauge what percentage of the email can be used for the calibration process. Since APWG does not tell whether an email is a Phishing, Spam, or Scam, we use the ScamPredictor [64] to determine an email as Phishing, Spam, or Scam.
- 34 emails are from a dataset we collected from other sources (i.e., universities, researchers, and tech companies). We use them because we have investigated them in the past and know what kinds of malicious emails they are. Among the 34, 18 are Phishing emails, 7 are Scam emails, and 9 are Spam emails. These 34 emails are only used in the *Calibration* process.

We characterize the emails that are respectively used in the *Calibration* and *Grading* processes as follows:

- Among the 141 (= 107 + 34) emails that are used in the *Calibration* process, 73 are used in the *Designing Grading Rules* sub-process and 68 are used in the *Training* sub-process. Among the 73 emails used in the *Designing Grading Rules* sub-process, 58.9% (43) are Phishing emails, 21.92% (16) Scam emails, and 19.18% (14) Spam emails. The temporal distribution is: 14 emails are from 2022; 8 emails from 2021; 14 emails from 2020; 9 emails from 2019; 2 emails from 2018; 1 email from 2017; 6 emails from 2016; 7 emails from 2015; 1 email each from 2014; 1 email from 2012; 3 emails from 2011; and 7 emails from 2006. Among the 68 emails used in the *Training* sub-process, 63.24% (or 43) are Phishing emails, 16.18% (or 11) are Scam emails, and 20.59% (or 14) are Spam emails. Their temporal distribution is: 15 emails are from 2023; 42 emails from 2022; and 11 from 2021. This uneven temporal distribution is caused by the fact that we use only the most recent emails so that we may piggyback them to help draw insights into the recent trend of malicious email sophistication.
- Among the 1,036 emails used in the *Grading* process, 64.86% (672) are Phishing emails, 21.53% (223) are Scam emails, and 13.61% (141) are Spam emails. The temporal distribution is: 496 emails are from the Decembers of 2006, 2011, 2016, and 2021, where 124 emails are randomly selected for each of these four months; 540 emails are selected September 1, 2021 and August 31, 2022, meaning that 45 emails are randomly selected from each of these 12 months, while noting that the 45 emails selected from December 2021 do not overlap with the 124 emails from the same month mentioned above. Note that we consider the Decembers of 2006, 2011, 2016, and 2021 because December is a holiday season that witnesses the highest malicious emails activities in a year [65], while noting that December 2021 is better represented with 169 (= 124 + 45) distinct emails. We select September 1, 2021-August 31, 2022 because it was the most recent year with available malicious emails from APWG at the time that we started this study. The random selection is achieved by using the Python NumPy and Pandas libraries. Note also that the date of an email is the date the attacker sent the email rather than the date the email was added to the APWG database.

Second, to assure that the malicious emails do not cause damage to the research environment, we use images to present them to the graders. The images are obtained by taking screenshots of the emails, while assuring that no link is clicked in the process.

Third, to assure that emails are reconstructed without missing elements (e.g., broken images), we make a `.eml` file based on the original raw email header and the body while incorporating the missing element. The emails are sanitized by removing their embedded warnings (e.g., "Warning!! This email comes from an external source!"), if present.

Fourth, displaying emails as images also assures that the emails are rendered the same over different computer platforms that may be used by the graders.

D. CALIBRATION

1) THE DESIGNING GRADING RULES SUB-PROCESS

This sub-process designs rules to guide the grading of emails with respect to PTechs and PTacs for determining agreement between graders and for resolving disagreements between graders. We also describe our grading aid.

a: RULES FOR GRADING WITH RESPECT TO PTechs

Grading rules with respect to PTechs are instructions on how to count the psychological elements corresponding to each PTech that is employed in an email. As mentioned above, we set the range for each PTech grade as $[0, 7] = \{0, 1, \ldots, 7\}$, namely $0 \le s_{i,j} \le 7$ for PTech$_i$.

- For Urgency, count the number of instances that trigger immediate action and force the recipient to act under time pressure (e.g., "now", "immediately", and "last chance").
- For Visual Deception, count the number of visuals that attempt to earn the trust of a recipient, such as logos.
- For Incentive & Motivator, count the number of instances of external rewards for incentives (e.g., "job offer", "50% discount"), and count the number of instances of internal rewards for motivators (e.g., "help me", "it's a patriotic thing to do");
- For Persuasion, count the number of the following principles [32] that are used in an email: (i) the AUTHORITY principle, by counting the number of instances of C-Suites titles (e.g., CEO, CFO), instances of providing expert opinion (e.g., from Dr. John Doe), and the number of real logos of known or trusted brands; (ii) the RECIPROCITY principle, by counting the

number of instances of demands for repayment of an earlier favor; (iii) the LIKING principle, by counting the number of instances of portrayal of similarities or common interests (e.g., referencing known people with same objectives); (iv) the SCARCITY principle, by counting the number of instances of any demonstration showing a lack or short supply of a goods or service; (v) the SOCIAL PROOF principle, by counting the presence of coercion (e.g.,"As an alumni of…"; (vi) the COMMITMENT principle, by counting the number of instances referencing earlier emails and/or the dedication to a cause/activity (e.g., "We are grateful for you past generosity"). Note that as justified above, this metric is upper bounded by 7 even if the actual count goes beyond 7.

- For **Impersonation**, count the number of pretense to be another entity (e.g., spoofed email address (domain & TLD), known personalities).
- For **Contextualization**, count the number of presences of contexts that are used to try to establish commonality with recipient, or mentioning relevant current events to engage the recipient (e.g., Covid-19).
- For **Personalization**, count the number of direct addresses of recipient (e.g., recipient's name, email address, telephone number).
- For **Attention Grabbing**, count the number of visual or auditory elements that prompt a recipient's focus (e.g., color button, bold font, upper case letters, highlighted text, bright-color font).

b: RULES FOR GRADING WITH RESPECT TO PTacs

Recall that for each PTac we set $\beta = 5$ as we use the rating scale $[0, 5]$ to measure the sophistication of an email with respect to a PTac. We propose using the following grading rule:

- '0' for *not applicable*, namely that a PTac is not employed in the email;
- '1' for *minimal* application, namely that the attacker does consider the PTac but applies it neither clearly nor consistently;
- '2' for *light* application of a PTac in an email, namely that the attacker considers the PTac, but with inconsistency, confusion, or lapses/errors in their approach;
- '3' for a *moderate* application of a PTac in an email, namely that the attacker clearly applies the PTac but may still have inconsistencies in their approach;
- '4' for a *significant* application of a PTac in an email, namely that the attacker clearly and consistently applies the PTac with minimal errors or lapses;
- '5' for an *extraordinary* application of a PTac in an email, namely that the attacker expertly and diligently crafts their email to apply the PTac in a cohesive and thoughtful way.

c: RULE FOR MEASURING THE DEGREE OF AGREEMENT BETWEEN GRADES

In order to measure the degree of agreement between the graders (i.e., their grades), we propose using the Kalpha metric (i.e., α) mentioned in the framework to measure the degree of agreement *between* the graders with respect to the PTechs or PTacs. This is reasonable because there could be many emails in question (e.g., 1,036 emails in our case study), meaning that we need to consider the *overall* agreement between the graders, and because the α is known to have the capability to capture the overall agreement. Nevertheless, we propose separating the treatment of the PTechs from that of the PTacs, namely that we will measure the degree of agreement among the graders with respect to the PTechs separately from that of the agreement among the graders with respect to the PTacs. This is reasonable because (i) the definition of email sophistication with respect to PTechs is separate from the definition of email sophistication with respect to PTacs and (ii) PTechs are measured using scale $[0, 7]$ while PTacs are measured using scale $[0, 5]$. In what follows we focus on the computation of the α with respect to PTechs, because its counterpart with respect to PTacs is treated in the same fashion.

For ease of reference, we denote $V = [0, 7]$ for PTechs. Recall in the framework that for a given email, we use $s_{i,j} \in V$ to denote the grade assigned to the email by grader j with respect to PTech i, where where $1 \leq i \leq \ell$ and $1 \leq j \leq n$. This is sufficient because we only need a method for measuring the degree of agreement among the PTech grades of all emails. In our case study, we need to use a specific method to measure the degree of agreement among all grades of all emails with respect to all PTechs. Since the specific method we employ needs to explicitly represent email identity k, where $1 \leq k \leq h$ and h is the number of emails in question, we extend the notation of $s_{i,j}$ to $s_{k,i,j}$, which is the grade of email $k \in [1, h]$ assigned by grader j with respect to PTech i. Note that the treatment of PTacs is the same, except that the number of PTacs is different from that of PTechs' and the scale or range V of PTacs is $V = [0, 5]$ (instead of $V = [0, 7]$).

Algorithm 1 is the method we advocate for computing the degree of agreement among the ℓ PTech grades of the h emails by all graders. The algorithm can be trivially adapted to compute the degree of agreement about the grades with respect to the PTacs. The algorithm is based on the method proposed in [17], which however does not give an algorithm description; this explains why we deem Algorithm 1 as a side product that may help computer scientists understand the method. The algorithm takes the $s_{k,i,j} \in V$'s as input and computes the α as the output.

The basic idea of the algorithm is the following. Lines 1-9 compute an agreement table T, or a $h \times \ell \times |V|$ matrix, where $V = \{0, 1, \ldots, 7\}$ in our case study and $T_{k,i,v}$ is the number of graders that assign the same grade $v \in V$ to email k with respect to PTech i. Using T, Lines 10-18

Algorithm 1 Computing Degree of Agreement α

```
INPUT: $\{s_{k,i,j}\}_{k\in[1,h],i\in[1,\ell],j\in[1,n]}$ where $s_{k,i,j} \in V$ is the grade of email $k$ in terms of PTech $i$ as assigned by grader $j$
OUTPUT: $\alpha$
1:  $T \leftarrow (T_{k,i,v} = 0)_{1\leq k\leq h, 1\leq i\leq \ell, v\in|V|}$ {initializing agreement table $T$}
2:  for $k = 1$ to $h$ do
3:      for $i = 1$ to $\ell$ do
4:          for $j = 1$ to $n$ do
5:              $v \leftarrow s_{k,i,j}$, where $s_{k,i,j} \in V$ {for better readability}
6:              $T_{k,i,v} \leftarrow T_{k,i,v} + 1$
7:          end for
8:      end for
9:  end for{$T_{k,i,v} = |\{j : s_{k,i,j} = v\}|$, namely the number of graders that assign grade $v$ to email $k$ in terms of PTech $i$, or (email, PTech)}
10: $\bar{T}_{k,i} \leftarrow 0$ for $1 \leq k \leq h$ and $1 \leq i \leq \ell${initialization}
11: for $k = 1$ to $h$ do
12:     for $i = 1$ to $\ell$ do
13:         for $v \in V$ do
14:             $\bar{T}_{k,i} \leftarrow \bar{T}_{k,i} + T_{k,i,v}$ {$\bar{T}_{k,i}$ is the number of valid graders with respect to a $(k, i)$ pair of (email, PTech)}
15:         end for{it is possible $\bar{T}_{k,i} < n$ as some $s_{k,i,j}$ may be invalid (i.e., outlier and not considered)}
16:     end for
17: end for
18: $\bar{T} \leftarrow \frac{1}{h\times\ell}\sum_{k=1}^{h}\sum_{i=1}^{\ell}\bar{T}_{k,i}$ {average number of valid graders per $(k, i)$ pair of (email, PTech)}
19: $\hat{T}_v \leftarrow 0$ for $v \in V${initialization}
20: for $v \in V$ do
21:     for $k = 1$ to $h$ do
22:         for $i = 1$ to $\ell$ do
23:             $\hat{T}_v \leftarrow \hat{T}_v + T_{k,i,v}$ {adding the number of valid graders per grade category $v \in V$ for all $(k, i)$ pairs of (email, PTech)}
24:         end for
25:     end for
26: end for
27: $\hat{T}_v \leftarrow \frac{\hat{T}_v}{h\ell}$ for $v \in V$ {average number of valid graders assigning grade $v \in V$ to the $(k, i)$ pairs of (email, PTech)}
28: $\pi_v \leftarrow \frac{\hat{T}_v}{\bar{T}}${probability a $(k, i)$ pair of (email, PTech) has grade $v$}
29: $T' \leftarrow T \times I_{|V|}$ where $I_{|V|}$ is $|V| \times |V| \times |V|$ identity matrix {see [66] for multiplication definition}
30: for $k = 1$ to $h$ do
31:     for $i = 1$ to $\ell$ do
32:         $p'_{a|(k,i)} \leftarrow \sum_{v\in V}\frac{T_{k,i,v}\left(T'_{k,i,v}-1\right)}{\bar{T}\left(\bar{T}_{k,i}-1\right)}$ {average ratio of agreement of graders assigning grade $v$ to a specific $(k, i)$ pair}
33:     end for
34: end for
35: $p'_a \leftarrow \frac{1}{h\times\ell}\sum_{k=1}^{h}\sum_{i=1}^{\ell}p'_{a|(k,i)}$ {average of ratio of agreement among all graders for all $(k, i)$ pairs}
36: $p_a \leftarrow \left(1 - \frac{1}{h\ell\bar{T}}\right)p'_a + \frac{1}{h\ell\bar{T}}$ {normalized average of agreement among all graders for all $(k, i)$ pairs}
37: $p_e \leftarrow \sum_{k=1}^{h}\sum_{i=1}^{\ell}T'\pi_v\pi_v^T$ {expected average of agreement when scoring is at random, where $^T$ means transpose}
38: return $\alpha \leftarrow \frac{p_a-p_e}{1-p_e}$
```

compute the average number of *valid* graders per (k, i) combination, namely per (email, PTech) combination, denoted by $\bar{T}$, where "valid" means that a grader's grade is not excluded; note that a grade can be excluded when it is an outlier, in which case the grade should be replaced with a value that does not belong to V (e.g., -1) so that the algorithm can be executed correctly. Line 27 computes the average number of graders that assign grade $v \in V$, denoted by $\hat{T}_v$. Line 28 computes the probability that a random grader assigns grade $v \in V$ to a (k, i), namely (email, PTech), pair, denoted by π_v. Line 29 computes an adjusted T based on the similarities among the grades [17], [67], denoted by T'. Lines 30-34 compute the average ratio of agreement for a (k, i) pair, denoted by $p'_{a|(k,i)}$. Line 35 computes the average of ratio of agreement among all the (k, i) pairs, denoted by p'_a. Line 36 computes the normalized average ratio of agreement, denoted by p_a. Line 37 computes the observed classification probability when grades are randomly assigned, denoted by p_e; this explains why α can offset the impact of random grades. Line 38 computes and returns the agreement α.

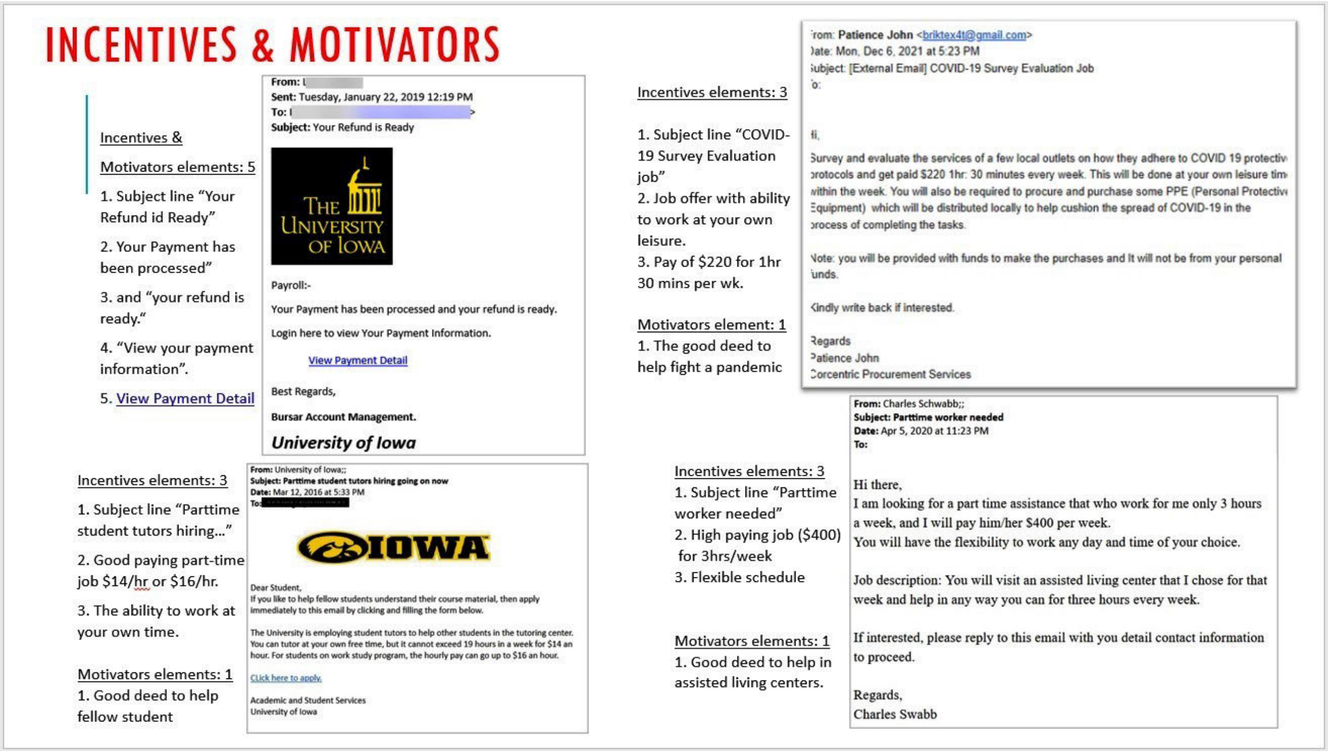
d: RULE FOR IDENTIFYING AND REMOVING / KEEPING OUTLIERS

In the case the grades are not in acceptable agreement with each other, we need to deal with outliers, which can happen in both the *Training* sub-process of the *Calibration* process and the *Analysis* process. For this purpose, we propose using the following heuristics to identify outliers, which are described with respect to PTechs but can be easily adapted to deal with PTacs. Note that there can be many approaches to dealing with outliers. We propose using the following approach because we have small number of graders ($n = 4$) and because the approach can reasonably handle the situations we encounter. We stress that these heuristics are not a standard statistical approach because n is small, and that social science studies do not necessary use standard statistical approach as demonstrated by the use of Kalpha mentioned above.

Recall that for a given email k and PTech i, the grades are $s_{i,j,k}$ for $1 \leq j \leq n$. Denote by $max_{i,k}$ the highest grade, namely $max_{i,k} = \max(\{s_{i,j,k}\}_{1\leq j\leq n})$; in the case of multiple grades are the highest, we choose an arbitrary one. Denote by $min_{i,k}$ the lowest grade, namely $min_{i,k} = \min(\{s_{i,j,k}\}_{1\leq j\leq n})$; in the case of multiple grades are the lowest, we choose an arbitrary one. We define the *spectrum* of the grades, denoted by $\delta_{i,k}$, as $\delta_{i,k} = max_{i,k} - min_{i,k}$. If $\delta_{i,k} < 3$, where '3' is chosen because for the scale $[0, 7]$ we have $\lfloor 7/2 \rfloor = 3$, we do not need to consider outliers. If $\delta_{i,k} \geq 3$, we still do not consider outliers under any of the following two scenarios (which stand out because we encountered them in our case study highlighted in Table 1): (i) the grade is a split, meaning that $\lfloor n/2 \rfloor$ graders assign one grade and the other $\lceil n/2 \rceil$ graders assign another grade; (ii) the grades are a sequence, such as $\{s_{i,j,k}\} = \{0, 1, 2, 3\}$. If $\delta_{i,k} > 3$ and none of the preceding two scenarios occurs, we need to determine which of $max_{i,k}$ and $min_{i,k}$ should be considered an outlier

TABLE 1. Grades of an email given by $n = 4$ graders, where grades with respect to the `Reward` PTac are a split and grades with respect to the `Fits & Form` PTac are a sequence (i.e., 0, 1, 2, 3); none of these grades should be eliminated as discussed in the text.

PTechs/PTacs		Grader 1	Grader 2	Grader 3	Grader 4	Min	Max	Spectrum	Std Dev
PTechs	Urgency	0	0	0	0	0	0	0	0
	Visual Deception	0	0	0	0	0	0	0	0
	Incentives & Motivators	1	0	1	2	0	2	2	0.71
	Persuasion	0	0	1	1	0	1	1	0.5
	Impersonation	2	1	2	1	1	2	1	0.5
	Contextualization	0	1	0	1	0	1	1	0.5
	Personalization	1	0	0	0	0	1	1	0.43
	Attention Grabbing	2	3	3	3	2	3	1	0.43
PTacs	`Familiarity`	1	2	3	2	1	3	2	0.71
	`Immediacy`	0	0	0	0	0	0	0	0
	`Reward`	0	0	3	3	0	3	3	1.5
	`Threat of Loss`	0	0	0	0	0	0	0	0
	`Threat to Identity`	0	0	0	0	0	0	0	0
	`Claim to legitimate Authority`	0	1	0	1	0	1	1	0.5
	`Fits and Form`	2	1	3	0	0	3	3	1.12



INCENTIVES & MOTIVATORS

Incentives & Motivators elements: 5

1. Subject line "Your Refund id Ready"
2. Your Payment has been processed"
3. and "your refund is ready."
4. "View your payment information".
5. View Payment Detail

From: [redacted]
Sent: Tuesday, January 22, 2019 12:19 PM
To: [redacted]
Subject: Your Refund is Ready

Payroll:-

Your Payment has been processed and your refund is ready.

Login here to view Your Payment Information.

View Payment Detail

Best Regards,

Bursar Account Management.

University of Iowa

Incentives elements: 3

1. Subject line "Parttime student tutors hiring..."
2. Good paying part-time job $14/hr or $16/hr.
3. The ability to work at your own time.

Motivators elements: 1

1. Good deed to help fellow student

From: University of Iowa;;
Subject: Parttime student tutors hiring going on now
Date: Mar 12, 2016 at 5:33 PM
To: [redacted]

Dear Student,
If you like to help fellow students understand their course material, then apply immediately to this email by clicking and filling the form below.

The University is employing student tutors to help other students in the tutoring center. You can tutor at your own free time, but it cannot exceed 19 hours in a week for $14 an hour. For students on work study program, the hourly pay can go up to $16 an hour.

Click here to apply.

Academic and Student Services
University of Iowa

Incentives elements: 3

1. Subject line "COVID-19 Survey Evaluation job"
2. Job offer with ability to work at your own leisure.
3. Pay of $220 for 1hr 30 mins per wk.

Motivators element: 1

1. The good deed to help fight a pandemic

rom: Patience John <briktex4t@gmail.com>
)ate: Mon, Dec 6, 2021 at 5:23 PM
ubject: [External Email] COVID-19 Survey Evaluation Job
o:

i,

Survey and evaluate the services of a few local outlets on how they adhere to COVID 19 protectiv protocols and get paid $220 1hr: 30 minutes every week. This will be done at your own leisure tim within the week. You will also be required to procure and purchase some PPE (Personal Protectiv Equipment) which will be distributed locally to help cushion the spread of COVID-19 in the process of completing the tasks.

Note: you will be provided with funds to make the purchases and It will not be from your personal unds.

Kindly write back if interested.

Regards
Patience John
Corcentric Procurement Services

Incentives elements: 3

1. Subject line "Parttime worker needed"
2. High paying job ($400) for 3hrs/week
3. Flexible schedule

Motivators elements: 1

1. Good deed to help in assisted living centers.

From: Charles Schwabb;;
Subject: Parttime worker needed
Date: Apr 5, 2020 at 11:23 PM
To:

Hi there,
I am looking for a part time assistance that who work for me only 3 hours a week, and I will pay him/her $400 per week.
You will have the flexibility to work any day and time of your choice.

Job description: You will visit an assisted living center that I chose for that week and help in any way you can for three hours every week.

If interested, please reply to this email with you detail contact information to proceed.

Regards,
Charles Swabb

FIGURE 3. Grading Aid Examples. The screenshot of a page in the grading aid document demonstrates how to count Incentives & Motivators cues in emails. For each email, it shows the different counts for cues of Incentives & Motivators, and which cues constitute incentives and which cues constitute motivators. The To field of the emails is redacted for privacy-protection purposes. Note the emails used in the grading aid are not used in the *Calibration* or *Grading* processes.

and eliminated. For this purpose, we propose comparing the following two distances: one distance is between the highest grade, $max_{i,k}$, and the second highest grade, denoted by max_2, which is the highest grade after removing max from the set $\{s_{i,j,k}\}$ (if there is a tie, we choose an arbitrary one); the other distance is between the lowest grade, $min_{i,k}$, and the second lowest grade, denoted by min_2, which is the lowest grade after removing min from the set $\{s_{i,j,k}\}$ (if there is a tie, we choose an arbitrary one). Then, we eliminate the one with a longer distance to its nearest neighbor; in the case of a tie, we do not eliminate any of them as outlier. Then, if $max - max_2 > min_2 - min$, then we treat max as outlier and eliminate it; if $max - max_2 < min_2 - min$, then we treat min as outlier and eliminate it; if If $max - max_2 = min_2 - min$, then we do not eliminate any of them.

e: DESIGNING AND DEVELOPING GRADING AID

Guided by the framework, we design and develop a grading aid, which is partly highlighted in Figure 3 to show how to count the number of instances of the Incentives & Motivators PTech.

TABLE 2. Example psychological elements of each PTech that can be used as cues exhibited in an email.

PTech	Example elements of PTech that may be exhibited in an email
Urgency	“call me now" / “Last chance to save your social life"' / “This is my last warning" / “Your PayPal access blocked!" / “I give your 72 hours to make the payment"
Visual Deception	PayPal logo, IRS logo / Replacing ‘womensright.com’ with ‘vvomensright.coom’ / Replacing ‘fbi.gov’ with ‘fbi.gov.net’; Replacing “microsoft.com" with micros0ft.com
Incentives & Motivators	“looking for a part-time assistant,(...) 3 hours a week, (...)$400 per week" / “Your refund notice" / “...get paid $220 1hr 30 minutes every week.
Persuasion	AUTHORITY - “...Common signs of infection include respiratory symptoms, fever, cough, breathing difficulties..."; COMMITMENT - “We are grateful for you past generosity"; LIKING “I’m sure you’ll agree with me"; RECIPROCITY - “We know you appreciate the scholarship you received ..."; SCARCITY - “We’ll send you a X hat which is limited in supply"; SOCIAL PROOF - “As an alumni of Calstatela, please consider joining... other alumni who have donated..."
Impersonation	“Yours sincerely, Warren Buffet" / “Dr. Eugene Gotcha, World Bank and Trust, 177a Bleecker Street, New York, NY 10012" / Phone/Fax number / CalStateLA Webmail Admin / “I’m the son of late millionaire president of Nigeria, Umaru Musa Yar’Adua”
Contextualization	“following the recent World Standing order over Corona Virus (Covid 19) pandemic..." / “Your UW.edu account..." / “The Fed cutting the interest rate to zero” / “Government emergency Covid-19 tax relief" / “...Ken State University retreat..."
Personalization	“Hi Wendy" / “Dear Jessica" / “Important message intended for John Doe" / “Your home at (home address) / “Your credit card ending in XXXX
Attention Grabbing	“CLICK HERE" / “Safety Measures.pdf" / “Important Covid-19 Updates & Measures" / “**Login here to action read**" / REVIEW NOW

- definitions of PTechs and PTacs, and a brief description of how they are employed in an email;
- examples of malicious emails describing how they are graded with respect to each PTech and PTac;
- an instantiated rating scale (i.e., $\beta = 5$) and how to grade the effectiveness of a PTac in an email (i.e., 0 means not applicable, 1 means minimal effort, 2 means light effort, 3 means moderate effort, 4 means significant effort, and 5 means extraordinary effort).
- a set of real-world emails used to show what constitutes a value on the instantiated scale [0, 5] for each PTac as mentioned above.

In addition, the grading aid also includes a reference table of key psychological elements associated with a PTech (e.g., the five examples of cues for the Urgency PTech and the five examples of cues for the Attention Grabbing PTech presented in Table 2, where colors and bold fonts are from real-world examples).

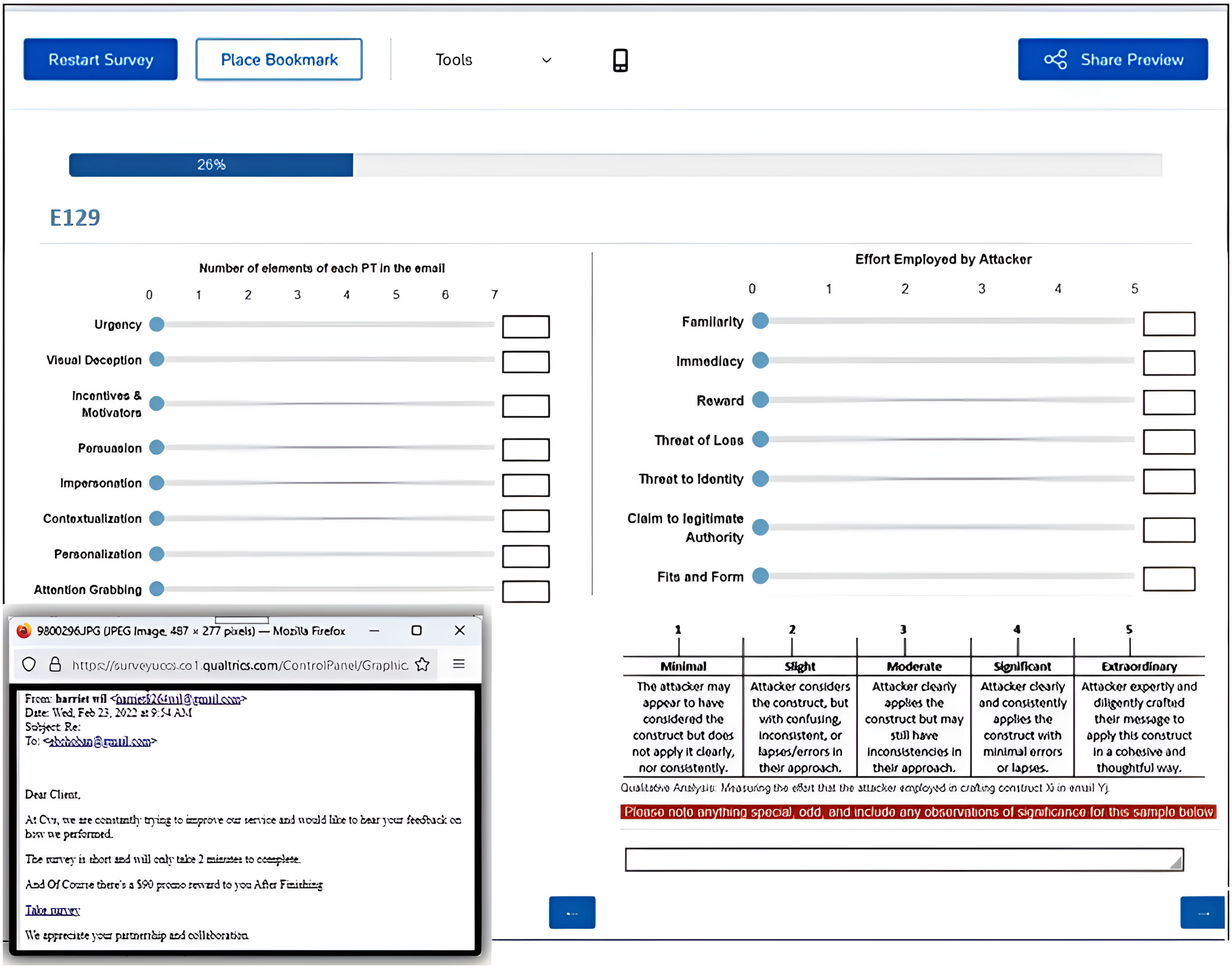
2) THE TRAINING SUB-PROCESS

Four cybersecurity PhD students with industry experience and cybersecurity certifications (e.g., CISSP), are recruited to follow the training process described in the framework. Among the four students, two have been publishing on cyber social engineering attacks through the psychological lens; the other two students have some knowledge in psychology based on their cybersecurity experience in coping with (e.g.) phishing attacks. The scheme was designed under the guidance of two professors who are experts in psychology. All four students understood the scheme well. The *training sub-process* includes the following steps below.

1) **Priming**. At this step, the graders learn the grading rules and the grading aid. The graders are also presented a series of email examples for the graders to assess and discuss as a group. The graders are primed on every item of the grading rules using sample emails. Questions are addressed, and opinions are reconciled.
2) **Testing**. At this step, each grader independently assesses a practice set of 52 emails (selected from the 141 emails mentioned above), while noting that $52/1,036 \approx 5\%$, which is deemed as an acceptable ratio. These 52 emails are made available in a survey format on the Qualtrics platform as shown in Figure 4. The α and standard deviation are computed based on the resulting grades.
3) **Evaluation**. The graders reconvene to discuss their grades and confirm a shared understanding of all grading rules. The α is used to determine the agreement among the graders. Then, the standard deviation is used to see the deviations between the grades, while identifying the PTechs and PTacs grades that exhibit more deviations than others.
4) **Resolution**. This step addresses and resolves the disagreements between the graders that are identified in the preceding step, as described in the framework.

E. GRADING

Our grading process addresses the issues raised in the framework as follows. First, as in the training process, the grading process is conducted using the Qualtrics platform. That is, both the emails and the grading aid are presented as pop-up windows. Second, to cope with the issue of grader’s fatigue, we divide the grading process into 11 self-paced sessions: nine sessions of 100 emails each, one session of 96 emails, and one session of 40 emails (for a total of 1,036). Third, to give the graders flexibility in the grading



1 Minimal	2 Slight	3 Moderate	4 Significant	5 Extraordinary
The attacker may appear to have considered the construct but does not apply it clearly, nor consistently.	Attacker considers the construct, but with confusing, inconsistent, or lapses/errors in their approach.	Attacker clearly applies the construct but may still have inconsistencies in their approach.	Attacker clearly and consistently applies the construct with minimal errors or lapses.	Attacker expertly and diligently crafted their message to apply this construct in a cohesive and thoughtful way.

Qualitative Analysis: Measuring the effort that the attacker employed in crafting construct Xi in email Yj

Please note anything special, odd, and include any observations of significance for this sample below

From: barriet wil <[illegible]@gmail.com>
Date: Wed, Feb 23, 2022 at 9:54 AM
Subject: Re:
To: <[illegible]@gmail.com>

Dear Client,

At Cvs, we are constantly trying to improve our service and would like to hear your feedback on how we performed.

The survey is short and will only take 2 minutes to complete.

And Of Course there's a $90 promo reward to you After Finishing

Take survey

We appreciate your partnership and collaboration.

FIGURE 4. A modified screenshot showing how a grader sees the survey on the Qualtrics platform. At the top there are options such as restarting the survey. A progress bar indicates the progress with respect to the number of emails for the session (usually 100 emails per session). E129 indicates the ID of the email that is currently being graded. The 8 PTechs are on the left side, and the 7 PTacs are on the right side, both having sliders and text boxes to input counts and grades respectively. The PTacs rating scale is at the bottom right (under the PTacs) to remind graders of what constitutes a grade from 1 to 5. The popup window (a separate window superimposed in this screenshot at the bottom left) portrays the screenshot of the email that is being graded. The popup screen changes to the next random email, but does not change the position of the window. A grader must grade all 8 PTechs and 7 PTacs before proceeding to the next email.

process, we ask each grader to conduct each session within 24 hours because we observe that each session of 100 emails may require about 5 hours on average to complete the grading. This gives a grader the flexibility to take breaks between grading sessions. Fourth, to mitigate the grading bias in terms of the order that emails are presented to graders, we randomize the order of emails within each session (i.e., different graders see emails in different orders).

An End-to-End Example of the Grading Process: To help understand how the grading process works, we use one email as as end-to-end example to demonstrate the process. The email is identified as E129 in our dataset and has an APWG ID# 116553-5424.24861 in the APWG database. In the *Preparing Data* process, we reconstruct the email from the raw data collected from APWG, convert the raw data into a `.eml` file, add the missing logo into the email, and remove warnings message in the email. Then, we open the `E129.eml` file using both Microsoft Outlook and Mozilla Thunderbird to make sure that all the components of the email are correctly displayed in both email clients. Finally, we take a screenshot of E129, and add the screenshot to the survey bin in the Qualtrics platform. The screenshot is displayed to the graders in the grading session as shown in Figure 4. Table 3 shows the respective grades assigned by the 4 graders to email E129. Results of all the emails is combined and analyze as an aggregate.

F. ANALYSIS

The 1,036 emails lead to 62,160 grades as each email is graded with respect to 8 PTechs and 7 PTacs and we

TABLE 3. Grades of an example email given by $n = 4$ graders. The Familiarity PTac has a spectrum $\delta_{i,k} = 4$, with standard deviation $\sigma = 1.5$. In this case, grade 5 is an outlier and eliminated according to the outlier removal rule, leading to a smaller spectrum $\delta_{i,k} = 1$ and smaller $\sigma = 0.47$.

PTechs/PTacs		Grader 1	Grader 2	Grader 3	Grader 4	Min	Max	Range	Std Dev
PTechs	Urgency	1	2	0	0	0	2	2	0.83
	Visual Deception	0	0	0	0	0	0	0	0
	Incentives & Motivators	1	1	3	2	1	3	2	0.83
	Persuasion	1	0	0	2	0	2	2	0.83
	Impersonation	2	3	4	4	2	4	2	0.83
	Contextualization	0	1	0	0	0	1	1	0.43
	Personalization	0	0	0	0	0	0	0	0
	Attention Grabbing	3	3	2	3	2	3	1	0.43
PTacs	Familiarity	1	2	2	5	1	5	4	1.50
	Immediacy	1	2	0	0	0	2	2	0.83
	Reward	2	2	2	2	2	2	0	0
	Threat of Loss	0	0	0	0	0	0	0	0
	Threat to Identity	0	0	0	0	0	0	0	0
	Claim to legitimate Authority	0	0	2	0	0	2	2	0.87
	Fits and Form	2	0	2	1	0	2	2	0.83

have 4 graders. Out of the 62,160 grades, there are 2,384 (i.e., 15.34%) outliers which are removed according to the outlier rule mentioned above corresponding to a PTech or PTac. Among the 2,384 outliers, 1,001 (3.02%) are out of the 33,152 PTech grades (i.e., 8 PTechs × 1,036 emails × 4 graders) and 1,383 (4.77%) are out of the 29,008 PTacs grades (i.e., 7 PTacs × 1,036 emails × 4 graders). We observe that 490 of the 1,036 emails had at least one outlier grade removed, and that split grades occur to 11 out of the 8,288 PTech grades (i.e., 8 PTechs × 1,036 emails) and 39 out of 7,252 PTac grades (i.e., 7 PTacs × 1,036 emails). In total, 1,001 outlier PTech grades are eliminated and 1,383 outlier PTac grades are eliminated. Before eliminating these outlier grades, we have $\alpha = 0.712$ for PTech grades and $\alpha = 0.605$ for PTac grades. According to the framework, we proceed to eliminate outliers, leading to 1,001 outlier PTech grades and 1,383 outlier PTac grades. After eliminating these outliers, we observe the α increased, with $\alpha = 0.822$ for PTech grades and $\alpha = 0.768$ for PTac grades, which are used because they represent very reliable agreement as described in the framework. The discrepancy between the agreement of the PTech grades (i.e., $\alpha = 0.822$) and the agreement of the PTac grades (i.e., $\alpha = 0.768$) can be attributed to the fact that PTac grades are more subjective, as shown by the fact that we eliminate more PTac outliers than PTech outliers.

Figure 5 plots the number of outliers with respect to each PTech and PTac. We observe the most outliers are incurred by the Familiarity PTac, the Impersonation PTech, and the Attention Grabbing PTech (in decreasing order), and that the least outliers are incurred by the Threat to Identity PTac, the Personalization PTech, and the Contextualization PTech (in increasing order). The implication is important: We need more training and research on the PTechs and PTacs so that graders can identify them more consistently. For graders, more consistent grading of malicious emails can better serve as the ground-truth in labeling the sophistication of malicious emails. For machine learning researchers, more training on PTechs and PTacs would allow them to more effectively define features to train machine learning models.

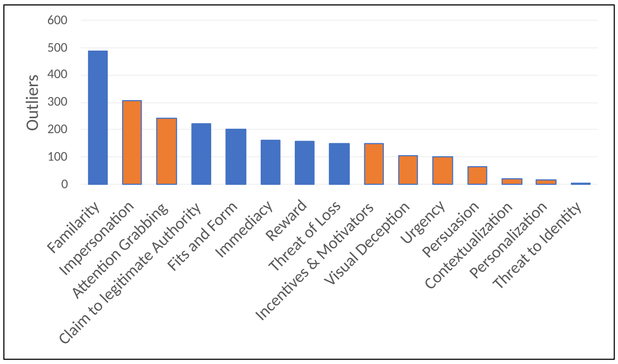


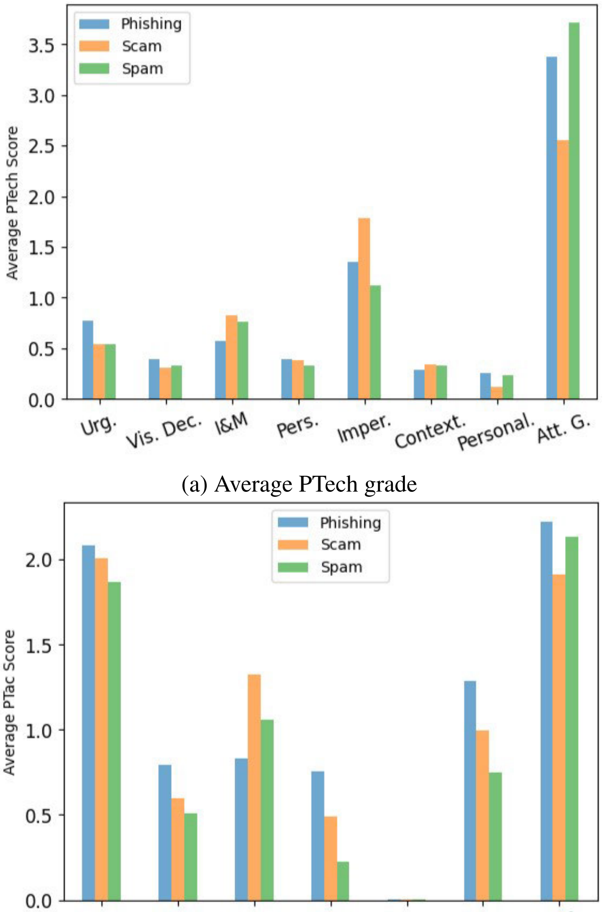
FIGURE 5. The total number of outliers for each PTech (orange color) and PTac (blue color).

Insight 1: To obtain ground-truth PTech and PTac sophistication, we need to provide further training and conduct further research so that graders can adequately approximate the ground-truth sophistication of malicious emails.

1) RQ 1: WHICH PTechs AND PTacs ARE WIDELY USED?

To identify the PTechs and PTacs that are most widely employed in each type of malicious emails, Figure 6a plots the average PTech grade, namely the average of the S_i's defined in Eq.(3) over the 1,036 emails, where for each PTech the score is the average of the *valid* grades (i.e., the grades remaining after eliminating outliers) over the number of *valid* graders whose grades are not eliminated as outliers. Figure 6b plots the average of the PTac grades, namely the average of the P_i's defined in Eq.(6) over the 1,036 emails, where the score for each PTac is the average of the valid grades over the number valid graders.

To see the overall use of PTechs and PTacs, Figure 7 plots the mean number of use of the 15 constructs across the three types of malicious emails. We make the following observations. First, Attention Grabbing is the most widely



(a) Average PTech grade

(b) Average PTac grade

FIGURE 6. Average PTech score (grade) and average PTac score (grade), where average is among the valid graders of the 1,036 emails.

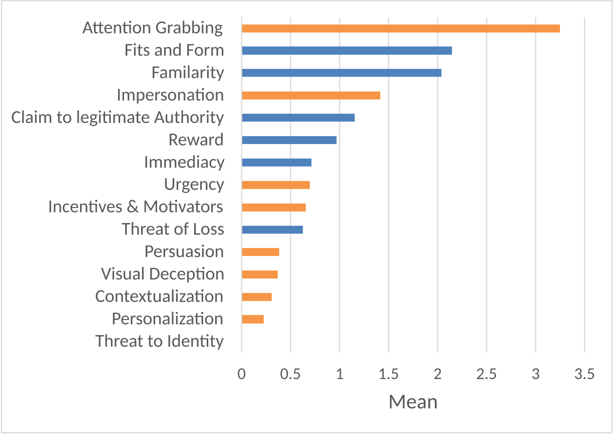


FIGURE 7. Mean value of all 15 constructs showing Attention Grabbing (PTech) with the highest mean value across all email types, followed by `Fit & Form` and `Familiarity` (PTacs).

employed PTech in all the three types of malicious emails, and is more widely employed in Spam emails than in Phishing and Scam emails. The Impersonation PTech is the second most employed PTech and distantly follows the Attention Grabbing PTech, and is more employed in Scam emails than in Phishing and Spam emails. The third most employed PTech does not follow the pattern of the first two most employed PTechs across all types of malicious email. While the Urgency PTech is the third most employed PTech in Phishing emails, the Incentives & Motivators PTech is the third most employed PTech in Scam and Spam emails. Similarly, the Incentives & Motivators PTech is the fourth most employed PTech in Phishing emails, but the Urgency PTech is the fourth most employed PTech in Scam and Spam emails. The least employed PTech is the Personalization PTech across all three types of malicious emails, perhaps because it is easier for attackers to send out a single generalized email to a large number of individuals than personalizing an email to a single individual. It is interesting to note that the Persuasion PTech, which is perhaps the most studied PTech in academic literature, is the fifth most widely employed PTech across the three types of malicious emails. This may explain why existing defenses are not as effective as desired because of the focus on the Persuasion PTech, which is not among the most widely employed PTechs.

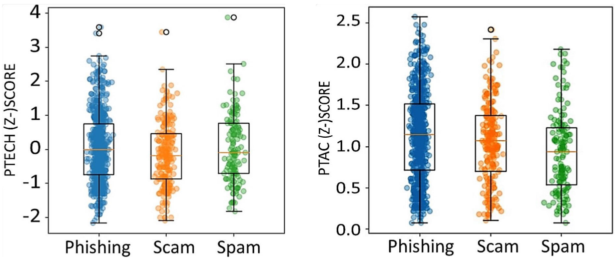
Second, the `Fit & Form` PTac is the most widely employed PTac in Phishing and Spam emails, but the second most employed PTac for Scam emails. It is closely followed by the `Familiarity` PTac, which is the most widely employed PTac in Scam emails. The `Reward` PTac is the third most employed PTac, but is most employed in Scam emails than in Spam and Phishing emails. `Threat to Identity` is the least employed PTac with an extraordinary low occurrence in all three types of malicious emails.

Insight 2: Existing studies might have focused on coping with the less proliferated PTechs (e.g., Persuasion*) and PTacs (e.g.,* `Reward`*), rather than the most proliferated PTechs (e.g.,* Attention Grabbing *and* Impersonation*) and PTacs (e.g.,* `Fit & Form` *and* `Familiarity`*).*

2) RQ 2: WHAT EMAILS ARE MORE SOPHISTICATED?

To understand how PTechs and PTacs may differ among the three types of malicious emails, we use the z-score method [68] to normalize the PTech grades to a scale comparable to the PTac grades. The z-score replaces a value by subtracting the mean value from it; dividing the result by the standard deviation. The mean values of the normalized PTech grades (i.e., 0.0425 for Phishing emails, 0.0258 for Scam emails, and -0.1443 for Spam emails) are more diverse than the mean values of the PTac grades, which concentrate in a small interval (i.e., 1.1418 for Phishing, 0.9351 for Scam, and 1.0472 for Spam). This difference suggests that PTech and PTac capture different aspects of malicious emails (i.e., both are important).

Figure 8a further shows an almost equal z-score between Phishing and Spams above Scams, and Figure 8b shows that Phishing emails exhibit a higher z-score that Scam and Spam emails. Cognizant of the fact that the sophistication of a malicious email is a two-dimensional vector (S_{PTech}, S_{PTac}),



(a) PTech grades. (b) PTac grades.

FIGURE 8. Boxplots of the normalized PTech grades and the original PTac grades.

where S_{PTac} measures the overall coherence and quality of the message, a high PTac z-score for Phishing coupled with a high PTech score makes Phishing emails more sophisticated than the other two kinds of malicious emails. This leads to:

Insight 3: Phishing emails are psychologically more sophisticated than Spam and Scam emails.

3) RQ 3: HOW ARE PTECHS AND PTACS CORRELATED?

To answer this RQ, Figure 9 summarizes the Pearson coefficient [18], which measures the correlation between the PTech grades, between the PTac grades, and between the PTech grades and the PTac grades.

a: PTech-PTech CORRELATIONS

The correlation between the PTechs is weak, with about an equal distribution of negative and positive correlation coefficients falling in between −0.2 and 0.3. The highest positive correlation among the PTechs is between Attention Grabbing and Visual Deception with a coefficient of 0.337, perhaps because Visual Deception requires some elements of Attention Grabbing to be effective (i.e., attackers may use visual elements to draw recipient's attention, such as displaying Dropbox logo with a button beneath that reads **Click here to download**, while the URL actually redirects to a malicious website rather than the Dropbox intended by the recipient). The second and third highest positive correlations are between Persuasion and Impersonation (0.244) and between Persuasion and Visual Deception (0.202), respectively. There are also negative correlations between the PTechs, with the highest negative correlation being between Impersonation and Attention Grabbing, with a −0.202 coefficient of correlation. This may be because Impersonation usually mimics an entity that does not use the elements of Attention Grabbing in their emails. The remaining negative correlations are too weak.

Insight 4: There is no strong correlation between the PTechs, suggesting that they are relatively independent of each other.

b: PTac-PTac CORRELATIONS

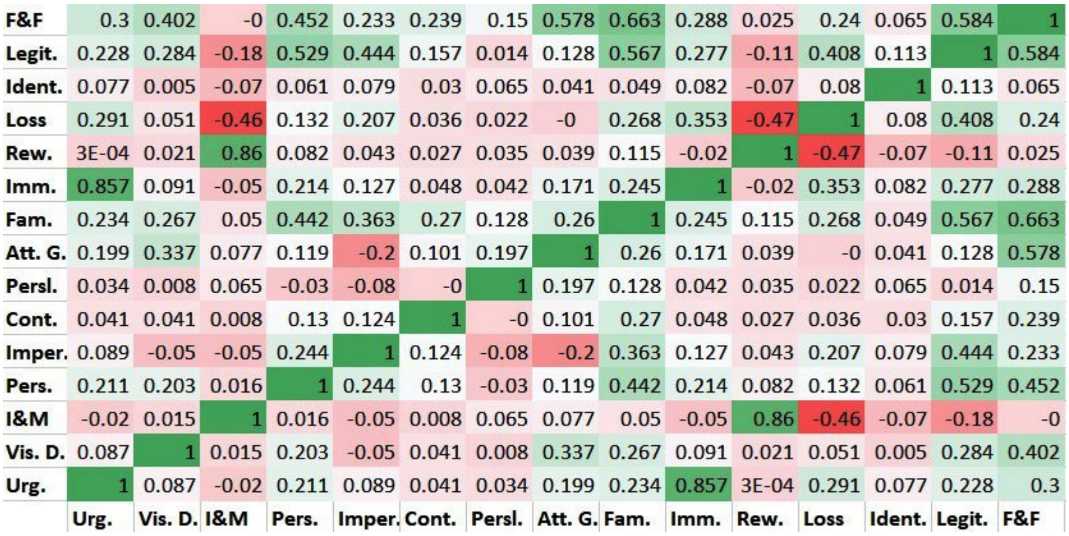
The correlations between PTacs are generally stronger than the correlations between the PTech. The distribution of the correlations is between −0.466 and 0.663. The highest positive correlation is between `Familiarity` and `Fit & Form` with a correlation coefficient of 0.663. This correlation may be due to the fact that attackers want the recipient of the email to lack suspicion about the email. Therefore, they make an email to fit the recipient's expectation of how such an email should look and feel. The second highest positive correlation among PTacs is between `Fit & Form` and `Claim to Legitimate Authority`, with a coefficient of 0.584. This may be due to the fact that attackers need to frame the email to fit the expectation of the authority that the attacker claims in order to curtail the suspicion of the email recipient. The third highest positive correlation among PTacs is between `Familiarity` and `Claim to Legitimate Authority`, with a coefficient of 0.567. This can be explained by the fact that claiming an authority is worthless if the recipient is not familiar with that authority. Therefore, to succeed in their attacks, attackers tend to claim authorities that are familiar to the email recipients.

The highest negative correlation among PTacs is between `Reward` and `Threat of Loss`, with a correlation coefficient of −0.466. This may be due to the fact that using threats to present a reward has the opposite effect; the greater the threat, the lesser the intended reward. It may also be due to the fact that presenting a reward through threat is counter-intuitive. Therefore, when one of these two PTacs is employed in an email, the other is absent, especially since a reward has to be softly presented as a bait rather than a threat.

Insight 5: `Claim to Legitimate Authority` *is the most correlated PTac with other PTacs, probably because attackers tend to claim a legitimate authority that is familiar to the email recipient together with other PTacs.*

c: PTech-PTac CORRELATIONS

The highest positive correlation between PTechs and PTacs is between the Incentive & Motivator PTech and the `Reward` PTac, with a correlation coefficient of 0.860. This is very closely followed by the correlation between the Urgency PTech and the `Immediacy` PTac, with a correlation coefficient of 0.857. The former correlation may be due to the fact that attackers usually present tangible goods such as money (i.e., a reward for doing something) or "free stuff," or "help others in need" in order to incentivize or motivate a recipient to take action. The latter may be due to the fact that both the Urgency PTech and the `Immediacy` PTac involve time, and whenever the Urgency PTech is employed in an email, it triggers the employment of the `Immediacy` PTac. The third highest correlation is between the Attention Grabbing PTech and the `Fit & Form` PTac, with a correlation coefficient of 0.578. There is also a strong positive correlations between the Persuasion PTech and the `Claim to Legitimate Authority` PTac (0.529), and between the Persuasion PTech and the `Fit & Form` PTac (0.4525). The highest correlation is between the Incentive & Motivator PTech and the `Reward`



F&F	0.3	0.402	-0	0.452	0.233	0.239	0.15	0.578	0.663	0.288	0.025	0.24	0.065	0.584	1
Legit.	0.228	0.284	-0.18	0.529	0.444	0.157	0.014	0.128	0.567	0.277	-0.11	0.408	0.113	1	0.584
Ident.	0.077	0.005	-0.07	0.061	0.079	0.03	0.065	0.041	0.049	0.082	-0.07	0.08	1	0.113	0.065
Loss	0.291	0.051	-0.46	0.132	0.207	0.036	0.022	-0	0.268	0.353	-0.47	1	0.08	0.408	0.24
Rew.	3E-04	0.021	0.86	0.082	0.043	0.027	0.035	0.039	0.115	-0.02	1	-0.47	-0.07	-0.11	0.025
Imm.	0.857	0.091	-0.05	0.214	0.127	0.048	0.042	0.171	0.245	1	-0.02	0.353	0.082	0.277	0.288
Fam.	0.234	0.267	0.05	0.442	0.363	0.27	0.128	0.26	1	0.245	0.115	0.268	0.049	0.567	0.663
Att. G.	0.199	0.337	0.077	0.119	-0.2	0.101	0.197	1	0.26	0.171	0.039	-0	0.041	0.128	0.578
Persl.	0.034	0.008	0.065	-0.03	-0.08	-0	1	0.197	0.128	0.042	0.035	0.022	0.065	0.014	0.15
Cont.	0.041	0.041	0.008	0.13	0.124	1	-0	0.101	0.27	0.048	0.027	0.036	0.03	0.157	0.239
Imper.	0.089	-0.05	-0.05	0.244	1	0.124	-0.08	-0.2	0.363	0.127	0.043	0.207	0.079	0.444	0.233
Pers.	0.211	0.203	0.016	1	0.244	0.13	-0.03	0.119	0.442	0.214	0.082	0.132	0.061	0.529	0.452
I&M	-0.02	0.015	1	0.016	-0.05	0.008	0.065	0.077	0.05	-0.05	0.86	-0.46	-0.07	-0.18	-0
Vis. D.	0.087	1	0.015	0.203	-0.05	0.041	0.008	0.337	0.267	0.091	0.021	0.051	0.005	0.284	0.402
Urg.	1	0.087	-0.02	0.211	0.089	0.041	0.034	0.199	0.234	0.857	3E-04	0.291	0.077	0.228	0.3
	Urg.	Vis. D.	I&M	Pers.	Imper.	Cont.	Persl.	Att. G.	Fam.	Imm.	Rew.	Loss	Ident.	Legit.	F&F

FIGURE 9. Correlations between the PTechs and PTacs presented in both values and colors. The darker the color (i.e., green for positive and red for negative) the stronger the correlation. The diagonal green line indicates each PTech and PTac has a 100% correlation with itself. The very small negative correlation values are seen as −0. The abbreviations are Urgency (Urg.), Visual Deception (Vis. D.), Incentives & Motivators (I&M), Persuasion (Pers.), Impersonation (Imper.), Contextualization (Cont.), Personalization (Persl.), Attention Grabbing (Att. G.), Familiarity **(Fam.),** Immediacy **(Imm.),** Reward **(Rew.),** Threat of Loss **(Loss),** Threat to Identity **(Ident.),** Claim to legitimate Authority **(Legit.),** Fits and Form **(F&F).**

PTac, closely followed by correlation between the Urgency PTech and the Immediacy PTac.

The highest negative correlation is between the Incentive & Motivator PTech and the Threat of Loss PTac, with a correlation coefficient of −0.464. This may be due to the fact that an incentive may not achieve its goal if it is presented in an email using threats. This may also be due to human nature, that a higher degree of threat triggers a recipient's higher degree of analytic reasoning and thus a higher degree of detecting that the email is malicious. Therefore, attackers avoid using threat in an email if they want to motivate recipients to respond positively to the attackers' requests. The second highest negative correlations is between the Incentive & Motivator PTech and the Claim to Legitimate Authority PTac with a correlation coefficient of −0.183, which is weak correlation. This may be because such emails mostly portray the power of authority rather than the skill to motivate. The PTech-PTac correlation shows that less sophisticated emails from the PTech perspective are also less sophisticated from the PTac perspective.

Insight 6: There is a strong correlation between the employment of PTechs and PTacs in malicious emails.

4) RQ 4: WHY EMAILS ARE INCONSISTENTLY GRADED?

To answer this RQ we reexamine two outcomes from grading the emails: (i) emails with standard deviations ($\sigma \geq 2$); and (ii) emails with split grades.

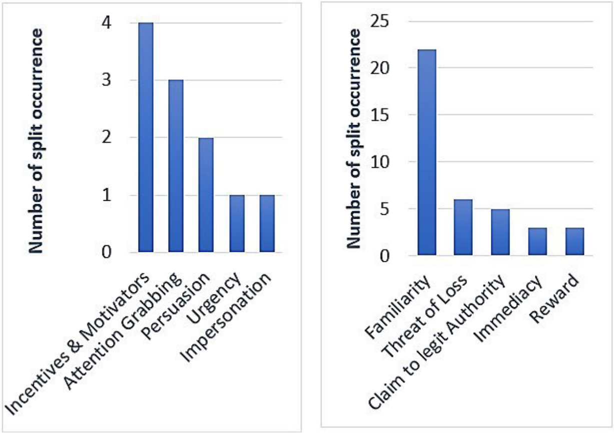
a: EMAILS WITH HIGH STANDARD DEVIATIONS ($\sigma \geq 2$)

There are 159 emails whose PTech or PTac grades have a $\sigma \geq 2$. A further analysis of these emails shows that they are highly charged with the Impersonation PTech, with 54.72% (87 out of 159) occurrences. The Impersonation PTech is to make a recipient think the attacker is someone they know. Therefore, this PTech can skew an email to be high or low depending on the grader's familiarity with the persona that the attacker assumes in the email. This large standard deviation can make an email to be graded high or low depending on: (i) the persona that the attacker assumes; and (ii) the familiarity of the grader with the persona. This can result in an email graded high by some graders and low by other graders. Note also that the persona the attacker assumes can also be an entity such as a known company or any such group.

b: EMAILS WITH SPLIT GRADES

We reexamine the emails with split grades. Out of 50 split grades occurring to 49 emails (i.e., one email had two split grades), 11 split grades occur to PTechs and 39 occur to PTacs. Figure 10 plots the number of split occurrences in PTech and PTac grades. We observe that 5 (out of the 8) PTechs, as indicated in the x-axis, has at least one split grade, and the Incentive & Motivator PTech has 4 (out of the 11) split grades. We observe that 5 (out of the 7) PTacs, as shown in the x-axis, has at least one split grade, among which Familiarity has 22 (out of the 39) split grades and the Threat of Loss PTac has 6 (out of the 39) split grades.

In terms of split grades per email type, Table 4 summarizes the number of split grades per PTech and per PTac. We observe that Phishing emails cause more split grades, including 8 (out of the 11) split PTech grades corresponding to 11 emails, and 23 (out of the 29) split PTac grades with respect to 28 emails. Since split grades occur only with some PTechs (i.e., 5 out of 8 PTechs) and PTacs (i.e., 4 out



(a) Split PTech grades. (b) Split PTac grades.

FIGURE 10. Split grade occurrences for PTechs and PTacs, where the *x*-axis shows the PTech or PTac that has at least one split grade.

TABLE 4. Summary of split grades with respect to PTechs and PTacs, per email type, where the abbreviations are Visual Deception (Visual Decept.), Incentives & Motivators (I&M), Contextualization (Contextualiz.), Attention Grabbing (Attention Grab.), `Threat of Loss (Loss)`, `Threat to Identity (Identity)`, `Claim to legitimate Authority (Legit Auth)`, `Fits and Form (F&F)`.

PTechs/PTacs		Phishing	Scam	Spam	Total
PTechs	Urgency	1			1
	Visual Decept.				
	I & M	4			4
	Persuasion	1		1	2
	Impersonation			1	1
	Contextualiz.				
	Personalization				
	Attention Grab.	2	1		3
PTacs	`Familiarity`	15	4	3	22
	`Immediacy`		2	1	3
	`Reward`			3	3
	`Loss`	6			6
	`Identity`				
	`Legit Auth`	2	1	2	5
	`F& F`				
Total by Email type (%)		31 (62%)	8 (16%)	11 (22%)	50 (100%)

of 7 PTacs), the employment of these PTechs and PTacs in certain emails may have caused these emails to be graded high or low by different graders. It should be noted that the calibration process can reduce the subjectivity of the graders, but may not eliminate it. This also explains why `Familiarity` has the highest occurrence among all PTacs and PTechs; the content of an email may be familiar with one grader but not with another. This difference is also reflected by the fact that an email may be graded high by one grader familiar with the email content, and low by another grader not familiar with it.

Insight 7: Malicious emails with content familiar to a recipient have a higher sophistication.

5) RQ 5: DOES SOPHISTICATION EVOLVE WITH TIME?

Figure 11 plots the yearly mean of the PTech grades and of the PTac grades using the aforementioned 124 emails in years 2006, 2011, 2016, and 2021, without eliminating any outlier (if applicable). We make two observations.

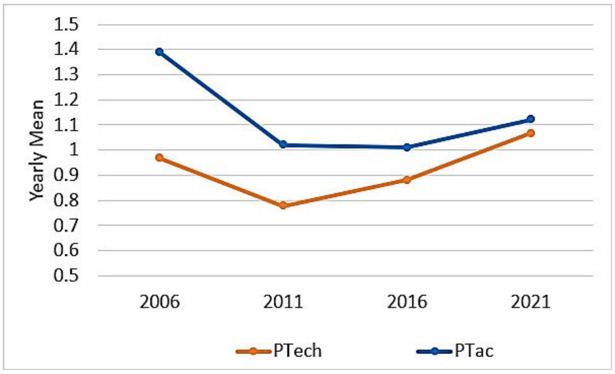


FIGURE 11. The average PTech and PTac grades (*y*-axis) in 2006, 2011, 2016, and 2021.

First, email sophistication decreases from 2006 to 2011, then monotonically increases from 2011 to 2021. By considering S_{PTech} and S_{PTac} separately, we observe: (i) the sophistication with respect to PTechs (or PTech-based sophistication) is higher in 2021 than in 2006, but the opposite is exhibited by that of the PTacs (or PTac-based sophistication); (ii) the difference between the PTech-based sophistication and the PTac-based sophistication has been decreasing, hinting that attackers have been making about the same efforts in improving their PTacs and PTechs.

Second, to see why there is drop in sophistication in 2011, we contrast the 124 emails in 2006 with the 124 emails in 2011 as well as their respective grades. We find that *email spoofing* (i.e., a malicious email uses a legitimate source address) and impersonation (e.g., the attacker impersonates another entity by using its name or logo) were more widely used in 2006 than in 2011. Specifically, 70% of the 124 emails in 2006 alleged to be coming from banks and financial institutions, while only 21% of the 124 emails in 2011 alleged to be coming from banks and financial institutions.

Third, to see why the PTac-based sophistication in 2006 is higher than that of 2021, we contrast the 124 emails in 2006 with the 124 emails in 2021 as well as their respective grades. We find that there is an increase in the diversification of impersonating entities. Most impersonating emails in 2006 alleged to be coming from 3 major institutions (i.e., Banks, PayPal, and Ebay); whereas, 2021 emails have impersonating entities that include banking and financial institutions, as well as delivery companies (e.g., USP, Fedex), Crypto platform (e.g., Coinbase), major stores (e.g., Kroger, Kohl's, Home Depot, Lowe's). The sophistication discrepancy between 2006 and 2021 could also be attributed to that defenses became effective against email spoofing and Impersonation around 2011, while these attacks were highly effective prior to 2011 as shown by the distribution of their message content over time [69]. As a result, the attackers turned to new techniques, which may not be as sophisticated as Impersonation but are effective, as existing defenses are yet to be effective against them [11].

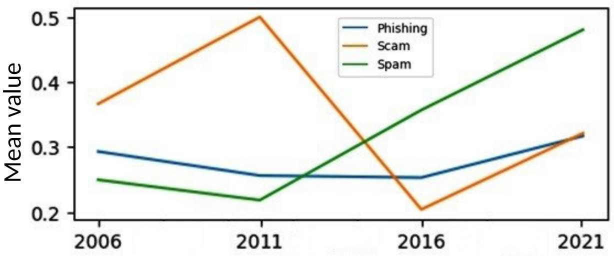


FIGURE 12. The average number of elements (i.e., cues) of the contextualization PTech (y-axis) exhibited by the 124 emails in 2006, 2011, 2016, and 2021, which reflects social events.

Insight 8: Effective defenses against PTechs and PTacs would force attackers to exploit other PTechs and PTacs. The trend from 2011 to 2021 indicates that attackers always improve their sophistication in both PTechs and PTacs over time.

6) RQ 6: HOW DO ATTACKERS EXPLOIT SOCIAL EVENTS?

We observe that all three email types leverage social events. We observe that Phishing emails leverage both happy events (e.g., Christmas and job offers) and sad events (e.g., flooding and pandemic); Scam emails leverage both happy events (e.g., new vaccine and mortgage refinancing) and sad events (e.g., earthquakes and wars); Spam emails leverage happy events (e.g., buying new homes and holidays sales) and sad events (e.g., health issues such as weight loss and pains). The exploitation of these events indicates that the attackers are *opportunistic attackers*, meaning that they leverage any event that can be used to wage attacks regardless of whether it is a tragic event or not.

Figure 12 plots the average contextualization, which reflects social events. We do not observe any patterns. Nevertheless, the significant increase of contextualization in Scam emails from 2006 to 2011 can be attributed to how attackers impersonated personalities with exorbitant wealth (e.g., Nigerian Prince, bank manager trying to exfiltrate unclaimed cash). Moreover, there seems to be templates for attackers to exploit social events, which can be evidenced by the fact that we observe 4 real-world emails that are essentially the same with the only difference that they impersonate 4 different companies (i.e., UPS, T-Mobile, Lowe's, and Home Depot).

Insight 9: Social events are widely exploited by attackers in their malicious emails.

V. LIMITATIONS

The present study has some limitations, which need to be addressed for the future. First, the framework has four limitations: (i) It is based on our understanding of psychological elements (i.e., PTechs and PTacs) that reflect the psychological sophistication of malicious emails. There may be other psychological elements that need to be considered, which can be accommodated by extending our framework. (ii) Its criteria for selecting PTechs and PTacs may not be perfect, meaning that the select PTechs and PTacs may not be complete. Nevertheless, the framework can be trivially extended to accommodate other PTechs and/or PTacs of interest, including those that may be introduced in the future. (iii) The grading rules may need to be refined, to more consistently ensure high levels of consistency in human assessment of email content. (iv) It advocates the use of Kalpha to measure the degree of consistency between graders. However, Kalpha, while widely used, may be less preferred by some researchers. In that case, they can replace the Kalpha with another method for measuring consistency of graders. For example, Kohen Kappa, which is a statistical measure of inter-rater reliability or how well raters agree on a variable, may be used when there are only two graders [70]. Fleiss Kappa, which is a statistical test that measures inter-rater reliability or the degree of agreement between multiple raters when assessing a categorical variable, may be used when there are more than two graders, where each grader only grades some but not all emails, or when the graders are randomly selected from a group of graders [57]. Unlike Kalpha, Fleiss Kappa does not cope with missing grades. Therefore, while other measures of inter-rater reliability can be used (depending on the number of graders), the number of emails, whether all graders grade all emails or not, or whether there are missing grades, we find Kalpha suitable for this study.

Second, the dataset has four limitations. (i) It may not be representative enough because we only collect and use emails from APWG, though it is arguably the most reputable source in the world. Moreover, even though manually grading 1,036 emails incurs a large amount of work load on each grader, the number of emails is considered small. Future studies need to seek automated grading methods to cope with much larger datasets. (ii) We admit the potential issue of 'informed' graders, as the graders are also the ones that design and revise the grading rules. This may affect the validity of the experimental results to some extent. Nevertheless, the fact that there are still many outliers suggests that the grading process is reliable because there would be no outliers otherwise. Still, future studies need to separate the group who design grading rules from the group who apply these rules to grade emails. (iii) While the framework can accommodate any reasonable definitions of PTech and PTac, it would be ideal to assure that the PTechs and PTacs are independent. For example, the Urgency PTech and the `Immediacy` PTac may overlap with each other, and future research needs to revise the PTechs and PTacs to make them independent of each other. (iv) Although the dataset is made up of malicious emails, legitimate emails may also employ PTechs. Given that we observed that Phishing emails are psychologically more sophisticated than Scam and Spam emails, it is reasonable to hypothesize that legitimate emails are not psychologically sophisticated. Regardless, the degree of sophistication of legitimate emails would not undermine the results of our study because one natural future study is to investigate the success of sophisticated malicious emails: Do more sophisticated emails necessarily have a higher success rate? (v) The current paper uses two key concepts: PTechs and PTacs. It is possible to extend the present study to accommodate the psychological factors (PFs) that are exploited by PTechs and PTacs, as described in [11].

VI. CONCLUSION

We have presented a framework for quantifying the psychological sophistication of malicious emails, including Phishing, Spam, and Scam emails. The framework is based on two aspects: PTechs and PTacs, both of which are necessary because they respectively reflect the low-level and high-level features of malicious emails. We defined metrics to quantify the sophistication of malicious emails. To measure these metrics, namely to approximate their ground-truth measurements as objectively as possible, we proposed grading rules to guide graders in measuring their sophistication with respect to the PTechs and PTacs. Based on a real-world dataset of 1,036 malicious emails and 4 graders, we draw a number of insights, which deepen our understanding of the sophistication of malicious emails and shed light on how to design effective defenses in the future.

There are interesting future research directions. (i) It is important to address the limitations of the present study mentioned above. (ii) It is important to extend the framework to cope with other types of cyber social engineering attacks (e.g., messaging-based attacks). (iii) Having showed that the concept of psychological sophistication is an inherent feature of malicious emails, it is important to investigate whether defenses, such as malicious email detectors, should be tailored to deal with malicious emails of different degrees of sophistication, or if we should seek "one size fits all" defensive mechanism (i.e., one mechanism that is effective regardless of the degree of sophistication of malicious emails). (iv) One approach to leveraging the concept of psychological sophistication to guide the design of effective defense is to design a tool that can automatically quantify the sophistication of an incoming email. This tool would need to be supported by multiple capabilities, such as: (a) automatically recognizing the images contained in emails and then automatically grading their sophistication according to the PTechs; (b) automatically grading the sophistication of incoming emails with respect to PTacs, which would require comprehension of the overall content of an email; and (c) showing that benign emails exhibit different sophistication characteristics (e.g., less sophisticated) than malicious emails, while noting that our framework is equally applicable to quantify the sophistication of benign emails.

ACKNOWLEDGMENT

The authors would like to thank the APWG for providing most of the emails that are used in the present study. Approved for Public Release; Distribution Unlimited. Public Release Case Number 23-3230. Rosana Montañez Rodriguez is also affiliated with The MITRE Corporation, which is provided for identification purposes only and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by them.

REFERENCES

[1] (2021). *Data Breach Investigation Report*.

[2] M. M. A. Pritom and S. Xu, "Supporting law-enforcement to cope with blacklisted websites: Framework and case study," in *Proc. 10th IEEE Conf. Commun. Netw. Secur.*, 2022, pp. 181–189.

[3] M. M. Ahsan Pritom, K. M. Schweitzer, R. M. Bateman, M. Xu, and S. Xu, "Characterizing the landscape of COVID-19 themed cyberattacks and defenses," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Nov. 2020, pp. 1–6.

[4] M. M. Ahsan Pritom, K. M. Schweitzer, R. M. Bateman, M. Xu, and S. Xu, "Data-driven characterization and detection of COVID-19 themed malicious websites," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Nov. 2020, pp. 1–6.

[5] T. Nelms, R. Perdisci, M. Antonakakis, and M. Ahamad, "Towards measuring and mitigating social engineering software download attacks," in *Proc. 25th USENIX Secur. Symp.*, Sep. 2016, pp. 773–789.

[6] M. Jakobsson, "The human factor in phishing," *Privacy Secur. Consum. Inf.*, vol. 7, no. 1, pp. 1–19, 2007.

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

[8] A. Vishwanath, T. Herath, R. Chen, J. Wang, and H. R. Rao, "Why do people get phished?" *Decis. Support Syst.*, vol. 51, no. 3, pp. 576–586, 2011.

[9] R. Anderson, *Security Engineering: A Guide To Building Dependable Distributed Systems*. Hoboken, NJ, USA: Wiley, 2020.

[10] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Apr. 2006, pp. 581–590.

[11] T. T. Longtchi, R. M. Rodriguez, L. Al-Shawaf, A. Atyabi, and S. Xu, "Internet-based social engineering psychology, attacks, and defenses: A survey," *Proc. IEEE*, vol. 112, no. 3, pp. 210–246, Mar. 2024.

[12] (2021). *Phishing Activity Trends Report-Unifying the Global Response To Cybercrime*.

[13] (2023). *Phishing Activity Trends Report-Unifying the Global Response To Cybercrime*.

[14] G. Ho, A. Cidon, L. Gavish, M. Schweighauser, V. Paxson, S. Savage, G. M. Voelker, and D. Wagner, "Detecting and characterizing lateral phishing at scale," in *Proc. 28th USENIX Secur. Symp.*, Aug. 2019, pp. 1273–1290.

[15] A. Oest, P. Zhang, B.Wardman, E. Nunes, J. Burgis, A. Zand, K. Thomas, A. Doupé, and G.-J. Ahn, "Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale," in *Proc. 29th Secur. Symp.*, Jan. 2020, pp. 361–377.

[16] H. Gascón, S. Ullrich, B. Stritter, and K. Rieck, "Reading between the lines: content-agnostic detection of spear-phishing emails," in *Proc. 21st Int. Symp.*, Jan. 2018, pp. 69–91.

[17] G. Marzi, M. Balzano, and D. Marchiori, "K-alpha calculator–Krippendorff's alpha calculator: A user-friendly tool for computing Krippendorff's alpha inter-rater reliability coefficient," *MethodsX*, vol. 12, Jun. 2024, Art. no. 102545.

[18] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia Analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.

[19] K. Krippendorff, "Computing krippendorff's alpha-reliability," *Empty*, vol. 2, pp. 1–20, May 2011.

[20] L. Allodi, T. Chotza, E. Panina, and N. Zannone, "The need for new antiphishing measures against spear-phishing attacks," *IEEE Secur. Privacy*, vol. 18, no. 2, pp. 23–34, Mar. 2020.

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

[22] A. v. d. Heijden and L. Allodi, "Cognitive triaging of phishing attacks," in *Proc. 28th USENIX Secur. Symp.*, May 2019, pp. 1309–1326.

[23] A. Ferreira and G. Lenzini, "An analysis of social engineering principles in effective phishing," in *Proc. Workshop Socio-Tech. Aspects Secur. Trust*, Jul. 2015, pp. 9–16.

[24] M. De Bona and F. Paci, "A real world study on employees' susceptibility to phishing attacks," in *Proc. 15th Int. Conf. Availability, Rel. Secur.*, Aug. 2020, pp. 1–10.

[25] L. Gallo, D. Gentile, S. Ruggiero, A. Botta, and G. Ventre, "The human factor in phishing: Collecting and analyzing user behavior when reading emails," *Comput. Secur.*, vol. 139, Apr. 2024, Art. no. 103671.

[26] L. Gallo, A. Maiello, A. Botta, and G. Ventre, "2 years in the anti-phishing group of a large company," *Comput. Secur.*, vol. 105, Jun. 2021, Art. no. 102259.

[27] A. Vance, D. Eargle, J. L. Jenkins, C. B. Kirwan, and B. B. Anderson, "The fog of warnings: How non-essential notifications blur with security warnings.," in *Proc. 15th Symp. Usable Privacy Secur. (SOUPS)*, Jan. 2019, pp. 407–420.

[28] S. Uebelacker and S. Quiel, "The social engineering personality framework," in *Proc. Workshop Socio-Tech. Aspects Secur. Trust*, Jul. 2014, pp. 24–30.

[29] Z. Wang, H. Zhu, and L. Sun, "Social engineering in cybersecurity: Effect mechanisms, human vulnerabilities and attack methods," *IEEE Access*, vol. 9, pp. 11895–11910, 2021.

[30] E. D. Frauenstein and S. Flowerday, "Susceptibility to phishing on social network sites: A personality information processing model," *Comput. Secur.*, vol. 94, Jul. 2020, Art. no. 101862.

[31] P. Schaab, K. Beckers, and S. Pape, "Social engineering defence mechanisms and counteracting training strategies," *Inf. Comput. Secur.*, vol. 25, no. 2, pp. 206–222, Jun. 2017.

[32] A. Ferreira, L. Coventry, and G. Lenzini, "Principles of persuasion in social engineering and their use in phishing," in *Proc. Int. Conf. Hum. Aspects Inf. Secur., Privacy, Trust*, Jan. 2015, pp. 36–47.

[33] J. Wang, T. Herath, R. Chen, A. Vishwanath, and H. R. Rao, "Research article phishing susceptibility: An investigation into the processing of a targeted spear phishing email," *IEEE Trans. Prof. Commun.*, vol. 55, no. 4, pp. 345–362, Dec. 2012.

[34] T. Longtchi and S. Xu, "Characterizing the evolution of psychological factors exploited by malicious emails," in *Proc. Int. Conf. Sci. Cyber Secur.*, 2024, pp. 1–26.

[35] T. Longtchi and S. Xu, "Characterizing the evolution of psychological tactics and techniques exploited by malicious emails," in *Proc. Int. Conf. Sci. Cyber Secur.*, 2024, pp. 1–12.

[36] (2023). *MITRE ATT&CK*. [Online]. Available: https://attack.mitre.org/

[37] R. M. Rodriguez and S. Xu, "Cyber social engineering kill chain," in *Proc. Sci. Cyber Security: 4th Int. Conf.*, Jan. 2022, pp. 487–504.

[38] N. H. Chowdhury, M. T. P. Adam, and G. Skinner, "The impact of time pressure on cybersecurity behaviour: A systematic literature review," *Behaviour Inf. Technol.*, vol. 38, no. 12, pp. 1290–1308, Dec. 2019.

[39] M. M. Moreno-Fernández, F. Blanco, P. Garaizar, and H. Matute, "Fishing for phishers. Improving Internet users' sensitivity to visual deception cues to prevent electronic fraud," *Comput. Hum. Behav.*, vol. 69, pp. 421–436, Apr. 2017.

[40] R. Montañez, E. Golob, and S. Xu, "Human cognition through the lens of social engineering cyberattacks," *Frontiers Psychol.*, vol. 11, p. 1755, Sep. 2020.

[41] H. Heckhausen, "Motivation as a function of expectancy and incentive," in *Motivation and Action*. Springer, 1991, pp. 113–161.

[42] F. Stajano and P. Wilson, "Understanding scam victims: Seven principles for systems security," *Commun. ACM*, vol. 54, no. 3, pp. 70–75, Mar. 2011.

[43] J. L. Freedman and S. C. Fraser, "Compliance without pressure: The foot-in-the-door technique.," *J. Personality Social Psychol.*, vol. 4, no. 2, pp. 195–202, 1966.

[44] M. Al-Hamar, R. Dawson, and L. Guan, "A culture of trust threatens security and privacy in Qatar," in *Proc. 10th IEEE Int. Conf. Comput. Inf. Technol.*, Jun. 2010, pp. 991–995.

[45] J. B. Hirsh, S. K. Kang, and G. V. Bodenhausen, "Personalized persuasion: Tailoring persuasive appeals to Recipients' personality traits," *Psychol. Sci.*, vol. 23, no. 6, pp. 578–581, Jun. 2012.

[46] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," *Commun. ACM*, vol. 50, no. 10, pp. 94–100, Sep. 2007.

[47] R. Montañez, A. Atyabi, and S. Xu, "Social engineering attacks and defenses in the physical world vs. Cyberspace: a contrast study," in *Cybersecurity and Cognitive Science*. Amsterdam, The Netherlands: Elsevier, 2022, pp. 3–41.

[48] T. Seitz, E. von Zezschwitz, S. Meitner, and H. Hussmann, "Influencing self-selected passwords through suggestions and the decoy effect," in *Proc. 1st Eur. Workshop Usable Secur.*, 2016, pp. 1–2.

[49] M. Cui, "How does the decoy effect affect decision-making and how we can prevent it?" in *Proc. Adv. Econ., Bus. Manage. Res.*, 2022, pp. 1753–1756.

[50] O. Gillath and G. Karantzas, "Attachment security priming: A systematic review," *Current Opinion Psychol.*, vol. 25, pp. 86–95, Feb. 2019.

[51] N. Benias and A. P. Markopoulos, "Hacking the human: Exploiting primordial instincts," in *Proc. South-Eastern Eur. Design Autom., Comput. Eng., Comput. Netw. Soc. Media Conf.*, Sep. 2018, pp. 1–6.

[52] S. C. Sethuraman, D. P. V S, T. Reddi, M. S. T. Reddy, and M. K. Khan, "A comprehensive examination of email spoofing: Issues and prospects for email security," *Comput. Secur.*, vol. 137, Feb. 2024, Art. no. 103600.

[53] P. Rajivan and C. Gonzalez, "Creative persuasion: A study on adversarial behaviors and strategies in phishing attacks," *Frontiers Psychol.*, vol. 9, p. 135, Feb. 2018.

[54] R. Wash, "How experts detect phishing scam emails," *Proc. ACM Human-Computer Interact.*, vol. 4, no. CSCW2, pp. 1–28, Oct. 2020.

[55] E. Ulqinaku, H. Assal, A. Abdou, S. Chiasson, and S. Čapkun, "Is real-time phishing eliminated with FIDO? Social engineering downgrade attacks against FIDO Protocols.," *IACR Cryptol. ePrint Arch.*, vol. 2020, p. 1298, Jan. 2020.

[56] F. Garcia-Loro, S. Martin, J. A. Ruipérez-Valiente, E. Sancristobal, and M. Castro, "Reviewing and analyzing peer review inter-rater reliability in a MOOC platform," *Comput. Educ.*, vol. 154, Sep. 2020, Art. no. 103894.

[57] A. Zapf, S. Castell, L. Morawietz, and A. Karch, "Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate?" *BMC Med. Res. Methodology*, vol. 16, no. 1, pp. 1–10, Dec. 2016.

[58] K. A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Tuts. Quant. Methods Psychol.*, vol. 8, no. 1, pp. 23–34, Feb. 2012.

[59] A. T. Jebb, V. Ng, and L. Tay, "A review of key Likert scale development advances: 1995–2019," *Frontiers Psychol.*, vol. 12, May 2021, Art. no. 637547.

[60] Anti-Phishing Working Group (APWG), 2024. https://apwg.org/

[61] (2024). *U.S.-CERT Federal Incident Notification Guidelines*. [Online]. Available: https://www.cisa.gov/sites/default/files/publications/

[62] (2024). *Where Does the 10% Condition Come From?*. [Online]. Available: https://web.ma.utexas.edu/users/mks/M358KInstr/TenPctCond.pdf

[63] K. Boddie. (2024). *Approximating Independence With the 10 Percent Rule*. [Online]. Available: https://study.com/skill/learn/approximating-independence-with-the-10-percent-rule-explanation.html

[64] (2023). *Scamdoc.com*. [Online]. Available: https://www.scamdoc.com/

[65] N. A. B. M. Zin, M. F. A. Razak, A. Firdaus, F. Ernawan, and N. S. A. Zulkifli, "Machine learning technique for phishing website detection," in *Proc. IEEE 8th Int. Conf. Softw. Eng. Comput. Syst. (ICSECS)*, vol. 17, Aug. 2023, pp. 235–239.

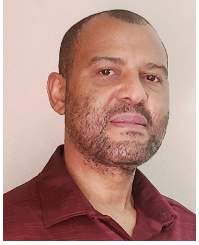
[66] C. Zaiontz. (2023). *Krippendorff's Alpha Basic Concepts*. [Online]. Available: https://real-statistics.com/reliability/interrater-reliability/krippendorffs-alpha/krippendorffs-alpha-basic-concepts/

[67] A. Mauboussin. (2024). *Inter-rater Reliability Metrics: An Introduction To Krippendorff's Alpha*. [Online]. Available: https://www.surgehq.ai//blog/inter-rater-reliability-metrics-an-introduction-to-krippendorffs-alpha

[68] T. Nield. (2022). *Essential Math for Data Science*.

[69] D. Wang, D. Irani, and C. Pu, "Is email bus. dying?: A study on evolution of email spam over fifteen years," *EAI Endorsed Trans. Collaborative Comput.*, vol. 1, no. 1, p. e3, May 2014.

[70] M. L. McHugh, "Interrater reliability: The Kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

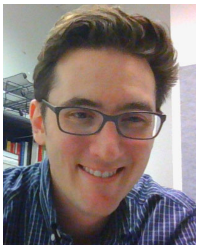


THEODORE TANGIE LONGTCHI received the B.S. degree in computer science and software engineering with concentration in information assurance and cybersecurity from the University of Washington, Washington, the M.S. degree in cybersecurity and leadership, and the Ph.D. degree in computer engineering with concentration in security from the University of Colorado Colorado Springs. His research interests include social engineering, cybersecurity, cognitive psychology, and the different domains of computer and information security. He has more than two decades of experience working directly or indirectly in different sectors of computer science and information technology while obtaining different certifications, such as the CompTIA Security+ and the Certified Information Systems Security Professional (CISSP).

ROSANA MONTAÑEZ RODRIGUEZ received the bachelor's degree in applied science from the University of Puerto Rico (Rio Piedras Campus), the bachelor's degree in computer science from the University of Maryland (University Campus), and the master's degree in security engineering from Southern Methodist University. She is currently pursuing the Ph.D. degree with the University of Texas at San Antonio. She works at The MITRE Corporation as a Cybersecurity Engineer. Her research explores the connection between human factors, cognitive psychology, and cybersecurity. Her research interests include cognition and security performance, computer-mediated information interpretation, and knowledge networks. She has more than 20 years of professional experience in IT operations and engineering, software engineering, and cybersecurity and has been a Certified Information Systems Security Professional (CISSP), since 2014.

KORA GWARTNEY received the Bachelor of Science degree in cybersecurity and digital forensics from National University, in 2019, and the Master of Professional Studies degree in cybersecurity analytics and operations from Penn State, in 2021. She is currently pursuing the Ph.D. degree with the University of Colorado Colorado Springs, under the advisement of Dr. Shouhuai Xu. She is the Lead Cybersecurity Engineer with Outside Interactive, specializing in security engineering. She is dedicated to exploring the intersection between humans and cybersecurity. She has 12 years of professional experience in the field and has a passion for understanding how human factors influence security with the aim to innovate and enhance cybersecurity practices through her research and professional work.

EKZHIN EAR received the Master of Science degree in cybersecurity, in 2019. He is currently pursuing the Ph.D. degree in security with the University of Colorado Colorado Springs, under the Department of Computer Science. He is a Lieutenant Colonel with U.S. Army, participating in its Advanced Civil Schooling program. He was commissioned as an Aviator through SROTC with Long Beach State University, in 2006, where he studied Computer Science. He transitioned to data systems engineering, in 2012, and has served in such positions as a System/Network Administer, an ISSM, a Senior Penetration Tester, a DCO Planner, and an OCO Analyst. He has held industry certifications from ISC2 (CISSP); Offensive Security (OSCE, OSCP); GIAC (GREM, GXPN, GPEN, GCPN, GWAPT); EC-Council (CEH, CHFI); CompTIA (Pentest+, Security+); CSA (CCSK); and Cisco (CCNA R&S, CCNA-S, CCNA-W).



DAVID P. AZARI received the B.S. degree in mathematics from the University of Colorado Boulder, in 2010, and the master's and Ph.D. degrees from the University of Wisconsin–Madison, in 2015 and 2018, respectively. He is currently an Assistant Professor of human factors and ergonomics within the Department of Behavioral Sciences and Leadership with the United States Air Force Academy, Colorado Springs, Colorado. He has interests in the intersection of human performance, simulation, and assessment in complex domains.

CHRISTOPHER P. KELLEY received the B.A. English and B.A. Sociology degrees from the University of Akron, in 2003 and 2004, respectively, and the M.A. degree in sociological social psychology and complex organizations and the Ph.D. degree from the University of Iowa, in 2008 and 2014, respectively. He is currently an Associate Professor with the Department of Behavioral Sciences and Leadership with the United States Air Force Academy, Colorado Springs, Colorado, where he is also the Sociology Directorate Lead. His research interests include basic social, interpersonal, and organizational processes.

SHOUHUAI XU (Senior Member, IEEE) received the Ph.D. degree in computer science from Fudan University. He is the Gallogly Chair Professor of cybersecurity with the Department of Computer Science, University of Colorado Colorado Springs (UCCS). He pioneered the cybersecurity dynamics framework as foundation for the emerging science of cybersecurity, with three pillars: first-principle cybersecurity modeling and analysis (the x-axis); cybersecurity data analytics (the y-axis); and cybersecurity metrics (the z-axis). The present study is integral to the cybersecurity dynamics framework because human factors cut across all these three pillars. He co-initiated the international conference on science of cyber security and is serving as its Steering Committee Chair. He is/was an Associate Editor of IEEE Transactions on Dependable and Secure Computing (IEEE TDSC), IEEE Transactions on Information Forensics and Security (IEEE T-IFS), IEEE Transactions on Network Science and Engineering (IEEE TNSE), and *Scientific Reports*. More information about his research can be found at https://xu-lab.org.