Evaluating the Impact of an Automated Email Analysis System on Phishing Susceptibility

Sean Strain

MInf Project (Part 2) Report
Master of Informatics
School of Informatics
University of Edinburgh

2023

Abstract

Users have been repeatedly shown to misinterpret the information within phishing emails that identify them as such, leading to incorrect assumptions of legitimacy. This paper and its predecessor outline the creation of a novel system that attempts to address this problem. This system evaluates information within emails that can conclusively identify phishing emails, and presents this to users automatically. In doing so, it may reduce the human-error involved in the user's evaluation process and thus improve their ability to detect phishing emails. Within this paper, a browser-based email inbox is modified to include this system. This simulated inbox is then used to conduct a quantitative study of 22 participants, aiming to evaluate how users interact with the system and the effect the system has on user phishing susceptibility. Users were found to have increased phishing detection precision and confidence. However, the study's limitations prevented further conclusions from being drawn. Nonetheless, this study motivates similar works in exploring the potential of automated user assistance in phishing detection.

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: 2022/82433.
Date when approval was obtained: 2022-11-20.
The participants' information sheet and a consent form are included in the appendix.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Sean Strain*)

Acknowledgements

While this paper and the work within is my own, I cannot ignore the contributions of my colleagues in the Technology Usability Lab in Privacy and Security (TULiPS) and at University College London. From supplying materials and software to guidance and advice, the work detailed within this paper and its predecessor simply would not be the same.

I would like to specifically mention Dr. Adam Jenkins, Dr. Nadin Kokciyan, Dr. Kami Vaniea, and Tarini Saka for their continuous help over the past two years. Further, I must thank Dr. Sarah Zheng for agreeing to supply this project with their source code for the website used within this study, and for the guidance in getting me up to speed on its workings.

This paper isn't just a thesis to me, it is also a way for me to transfer the work of the last two years of my life over to TULiPS. I hope that the team is able to continue on this work and build something that can really help people.

Table of Contents

1 Introduction **1**

2 Background and Relevant Work **3**
2.1 Email Phishing . . . 3
2.1.1 How Phishing Attacks Exploit Email and the User . . . 3
2.2 Quantifying Phishing Susceptibility . . . 5
2.2.1 User Motivations . . . 5
2.2.2 Understanding Why Users are Deceived by Phishing Attacks . 6
2.2.3 Methods of Quantifying User Susceptibility . . . 7
2.3 Summary . . . 9

3 Previous Work and Extensions **10**
3.1 Email Analysis Module . . . 10
3.2 Extensions Upon the Previous Work . . . 12
3.2.1 Additional Features . . . 12
3.2.2 Extensions . . . 13
3.2.3 Improvements . . . 14
3.3 Summary . . . 15

4 Methodology **17**
4.1 Overview . . . 17
4.2 Study Design . . . 18
4.2.1 Task Flow . . . 18
4.2.2 Experimental Conditions . . . 19
4.2.3 Email Set and Selection Methodology . . . 19
4.3 Analyses of Results . . . 21

5 Implementation **22**
5.1 Modifying the Survey Website . . . 22
5.1.1 Overview . . . 22
5.1.2 Usability . . . 23
5.1.3 Incorporating Email Analysis . . . 23
5.1.4 Changing the Email List . . . 24
5.1.5 Changes to User Flow . . . 26
5.1.6 Changes in Data Retrieval . . . 26
5.1.7 Summary . . . 27

5.2 Testing . . . 27
5.2.1 Think-Aloud Testing . . . 27
5.2.2 Web Browser Compatibility . . . 28
5.2.3 Software Testing . . . 29

6 User Study Results **30**
6.1 Exploratory Analysis of Module Usage . . . 30
6.2 User Performance . . . 31
6.2.1 Phishing Emails . . . 31
6.2.2 Legitimate Emails . . . 32
6.2.3 Module Impact on Inbox Processing Time . . . 34
6.3 Summary . . . 34

7 Discussion and Evaluation **35**
7.1 Interpretation of Results . . . 35
7.2 Limitations . . . 37
7.2.1 Participant Recruitment and Demographics . . . 37
7.2.2 Website Environment and Methodology . . . 37
7.2.3 Email Analysis Module . . . 38

8 Conclusion and Future Work **39**
8.1 Future Work . . . 39

Bibliography **41**

A Website Tests **52**
A.1 Think-Aloud Test and Results . . . 52
A.1.1 Participant One . . . 52
A.1.2 Participant Two . . . 52
A.2 Screenshot Tests . . . 53

B Study Website Design **55**
B.1 Unmodified Website . . . 55
B.2 Modified Website as Described in Section 5.1 . . . 58

C Participant Recruitment Email **68**

D Post-task Survey **70**

E Materials Employed in User Study **73**

F Microsoft Outlook Add-in **82**

G Generative Language Models in Phishing **85**

Chapter 1

Introduction

As the world has become increasingly digital, so too has the way we communicate with each other. One of the most prevalent forms of this digital communication - ubiquitous throughout institutions, workplaces, and academia worldwide - is that of electronic mail (email). Hundreds of billions of emails are sent every day between the 3.9 billion unique users possessing an email account [38]. Email is a global system allowing anyone to communicate with anyone. However, alongside technological progress comes the need to advance the methods we use to safeguard ourselves from malice; the ubiquity of email has allowed users to be accessible to deceptive schemes and exploitation.

The term 'phishing' refers to a class of social-engineering attacks wherein a fraudulent message is sent with the intent to deceive the recipient into believing the message is from a trustworthy source. Through exploiting the recipient's misplaced trust, the attacker intends to manipulate the recipient into engaging with the message in a way they would not if they knew the true sender. This could result in the recipient unwittingly visiting a malicious website [45], downloading malicious software [101], revealing sensitive information [67], etc. . 96% of phishing attacks are delivered via email [93, 116], constituting the most prevalent form of social-engineering attack [117] and the most common form of cybercrime [29, 80] worldwide.

The damage these attacks cause is significant. Estimates place the worldwide economic damage at millions of USD per *minute* [93]. This has motivated a range of research into both technical and user-focused solutions to prevent phishing attacks from succeeding.

For an email phishing attack to succeed the email must complete two things. First, it must bypass the automated mechanisms that detect and prevent malicious emails from reaching the recipient at all. Recent work has increased the detection rate of malicious emails upwards of 99.9% [57]. Despite this success, no system - nor any group of systems - will be able to guarantee that no phishing emails will ever reach their recipients [50]. The fact that phishing emails will arrive in user inboxes is inescapable. Thus, this leaves the final obstacle a phishing email must overcome: the human that reads it [59].

In 2022, 82% of corporate data breaches involved human error in some way, with phishing being the most common vector of attack [117]. Multiple studies have shown

that users often fail to actively engage themselves in evaluating the legitimacy of emails they receive [69, 78, 118]. Moreover, when they do, they often do not see [41], do not understand [26], or do not use [123] the evidence within emails that can conclusively identify them as phishing, leading to users reaching incorrect conclusions.

This presents two facts: phishing emails *will* arrive in user inboxes, and many of those users are susceptible to engaging with said emails in unsafe ways.

Motivated by these shortcomings, last year I proposed a system that addressed a gap in the existing literature [95]. I reviewed the common characteristics of phishing attacks, and discovered a wealth of information within emails that could be automatically extracted and evaluated to identify them. Thus, I developed a system - the email analysis module - that demonstrated that it is possible to automatically create a contextualised, human-readable report about any email's legitimacy using this information. This report attempts to alleviate the burden on the user by automatically evaluating the information they commonly misinterpret. Further, the process to create this report was measured in seconds as opposed to the tens of minutes it takes for an organisations' Security Operations Centre (SOC) to respond when a user asks for help on a particular email [32]. In my literature review I did not identify a previous system that performed these functions. I will describe this system in more detail in Chapter 3.

However, it was not demonstrated whether this novel system would be beneficial to a user's ability to identify phishing emails. If this is the case, the system may become a valuable addition to the phishing defence framework. Determining this is central to the work conducted in this paper. Thus, in this paper, I make the following contributions:

- Firstly, I will outline the technical and social vulnerabilities that exist within the email framework. I will then investigate the previous work conducted in the field of user-focused phishing literature; this will inform and motivate an appropriate approach to evaluating the aforementioned system.
- I will outline the work I completed in the previous year and the improvements made since. Subsequently, I will describe other systems I have created that build upon this work.
- I will describe the methodology behind a quantitative study of 22 participants that aimed to evaluate whether the information provided by the system assists users in identifying phishing emails, and to determine how the participants interact with the system.
- To conduct this study, I modified a website that simulates an email inbox to include the system. The process I undertook and challenges that arose are described.
- Finally, I will discuss the implications and limitations of said study before recommending a suggested direction for the future work.

This paper was conducted alongside the research of the Technology Usability Lab in Privacy and Security (TULiPS) at the University of Edinburgh (UoE), and builds upon the research conducted by Zhang in 2021 [130]. Zhang's work investigated potential features for the email analysis module, which I implemented last year and will evaluate in this paper.

Chapter 2

Background and Relevant Work

Within this chapter I will firstly outline the technical specifications of emails and the technical and social vulnerabilities that arise therein - a process conducted in more detail last year [95]. Subsequently, I will discuss relevant works within of the field of phishing research to motivate the work conducted within this paper.

2.1 Email Phishing

The technical specifications defining what an 'email' is are outlined across several documents and protocols [28, 108]. The current standardised email format is established in Request For Comments (RFC) 5322 [46, 90] and its extensions in the Multipurpose Internet Mail Extensions (MIME) standard [31]. This format mandates that an email must consist of two parts: the header and body.

The header is a structured set of fields containing information about the email. Such headers may include a *Message-ID* that uniquely identifies that email, or the *From* header that contains the sender's address. There are 184 potential headers an email may have [54]. The header precedes the body, which contains the information the sender intended for the recipient to see, which may be text, images, attached files, etc. .

Emails are sent/received by Mail User Agents (MUA) [108], also known as email clients, e.g., Microsoft Outlook [74]. MUAs use intermediary agents[1] when sending an email to another MUA, creating a 'received chain' of agents involved in transferring an email to the recipient. This process is defined in the Simple Mail Transfer Protocol (SMTP) [53]. Emails are accessed by recipients using the Internet Message Access Protocol (IMAP) [72]. Emails are predominantly stored using the 'de facto standard[]' [88] file type Electronic Mail Format (EML) [106].

2.1.1 How Phishing Attacks Exploit Email and the User

Within these specifications arise a number of technical limitations that can be exploited by malicious users [2, 15]. Further, there are a number of user-focused techniques an

[1] Mail Transfer Agents/Mail Delivery Agents.

email phishing attack may use to deceive a user [25, 26, 48, 56]. These include, but are not limited to, the following:

- **'Spoofing'** - SMTP alone does not define a way for the recipient to guarantee that the email arrived from the address it purports to be from [58]. In SMTP, anyone is able to send an email as if they were anyone else - a phenomenon known as 'spoofing'. To solve this, 4 additional authentication protocols have been created - Authenticated Received Chain (ARC), Sender Policy Framework (SPF), Domain Keys Identified Message (DKIM), and Domain-based Message Authentication, Reporting, and Conformance (DMARC) [22]. However, acceptance of these new standards - while increasing - is low [23]. Further, these headers may themselves be edited to appear as though the sender authenticated successfully[2] [112].
- **Email Addresses** - While spoofing allows for the exact replication of another address, attackers may purposefully give themselves an address that *closely resembles* other senders - a manipulation the aforementioned authentication protocols cannot solve [22] but may deceive users [26].
- **Attachments** - Attaching files to emails has legitimate uses, however 36% of phishing emails use attachments to send malware [17] that, if downloaded, may create malicious processes on the recipients' device [101].
- **Language** - Phishing attacks commonly feature predictable patterns in their use of language: heightened emotions [95], urgent language [2], poor grammar/spelling [2], not using the recipients' name [7], to using specific words often found in phishing emails [8].
- **Uniform Resource Locators (URLs)** - A URL is a reference to a networked file, document, web page, etc., which allow the user to access the referenced information [9]. These can be manipulated to allow a malicious sender to deceive recipients into believing the URL references a resource it does not [119]. The URL may reference a malicious resource, such as malware or a malicious website, while looking to many users as a reference to a legitimate resource [1, 119]. URLs are the most prevalent vector of attack in phishing emails [89], yet users often misunderstand and/or incorrectly evaluate the legitimacy of them [1, 25].
- **HTML** - the HyperText Markup Language (HTML) allows users to add colours, images, etc., to emails [124]. It also allows for the execution of JavaScript code [76], however, code execution is no longer tolerated in the vast majority of email clients [127]. A more relevant exploit is the HTML anchor (a) tag. This allows for further obfuscation of URLs in an email by making the text shown to the user (the tag's content) be an entirely different URL than the one it actually points to (the tag's Hypertext REFerence (href)) - a phenomenon I call *mismatching*.

[2] DKIM protects against this, but often the protocol may detect changes within the email without informing the recipient out side the email header.

2.2 Quantifying Phishing Susceptibility

Seeking to reach a scientific consensus upon the definition of phishing, Lastdrager defined the class of attacks as follows: “Phishing is a scalable act of deception whereby impersonation is used to obtain information from a target.” [59]. This target is not software, nor hardware, but the user [30]. As such, there is significant responsibility placed on the user to prevent the phishing attack from succeeding. While much research has investigated methods to alleviate this responsibility through technical solutions [3, 15, 50], research has also sought to understand why users are deceived by phishing attacks and how best to assist them in evaluating an emails legitimacy.

2.2.1 User Motivations

When attempting to understand a user’s susceptibility to phishing attacks, it must be acknowledged that preventing the success of phishing attacks is not the user’s primary goal [30, 125]. Herley remarks that the constant competition for users’ time and focus on the modern internet plays a significant role in shaping user priorities; security experts often do not provide sufficient evidence to convince users that investing time in security measures is worth more than their primary goals [42, 43]. Indeed, Herley asserts that users will typically exert only the minimum effort required when addressing security concerns [43]. A body of research supports this thinking, showing that users wilfully perform insecure behaviours because of the perceived benefit, be it in time or effort expended [44, 47, 125]. Thus, it is unreasonable to assume the user will always expend the increased effort required to check each detail of every email to evaluate legitimacy.

It is hypothesised that user motivations result in two forms of cognitive processing with regards to identifying phishing [69, 78, 118, 121]: the default, passive process requiring reduced user time and effort, and an active one where the user is more discerning and investing increased effort. The user may be prompted into the active state by various factors, such as cues within the email suggesting illegitimacy [122], explicit reminders of secure behaviour [113], or being informed that they will be assessed on their ability to identify phishing attempts [85], etc. .

A range of literature supports this thinking. Vishwanath et al. demonstrated that successful phishing attacks are often characterised by the deceived user having limited motivation, resulting in an insufficient examination of the email [118]. Luo et al. emphasise that the objective of a phishing attacker is to create messages that deter thorough processing, leading to quick, incorrect decisions [69]. Musuva et al. found that users with less phishing knowledge were less likely to expend effort in analysing an email, and that users who expended more effort analysing an email were more likely to correctly identify phishing [78]. Wash [122] and Wash et al. [123] showed that users passively read an email until they are triggered into a more active state of awareness by heuristic cues suggesting illegitimacy. Further, research in other areas of cybersecurity have identified that users need only be reminded of secure practices for them to engage more actively [113].

Parsons et al. sought to evaluate the effect of prompting users into a more active state on their ability to discern genuine emails from phishing [85]. They conducted a role-play

experiment with 117 participants managing 50 emails. The study investigated the impact of participants' awareness of the phishing study on their decisions. Half of the participants were informed about the phishing assessment. Those informed showed significantly better performance in identifying phishing emails. This increase was not because informed participants were biased towards judging an email as phishing, instead, they demonstrated increased ability to discriminate between genuine and phishing emails.

This effect, known as 'framing' or 'priming' [111], is considered deleterious to the generalisability of studies' results, as the results may not accurately reflect how users would respond to phishing attempts in real-world situations. Thus, many phishing studies have aimed to avoid prompting the user into an active state when measuring susceptibility [26, 55, 131]

However, Parsons et al.'s study and those previously discussed also asserts it is beneficial to engage the user when they are evaluating an email's legitimacy. Engaging the user in evaluating the email more thoroughly may prevent successful phishing attacks [118] and counteract phishing tactics that attempt to prevent such examination [69]. A study may build on this work by evaluating a system that engages the user when evaluating emails.

Such a study would first need to be informed by an understanding of why users are deceived by phishing attacks. Through this, it may be determined how the system could most effectively activate this higher level of processing.

2.2.2 Understanding Why Users are Deceived by Phishing Attacks

Several studies have explored the ways in which users engage with their emails and how they decide whether an email is suspicious. Through better understanding user behaviour, these studies inform future strategies in assisting users with detecting phishing attempts.

Wash explored how experts approach suspicious emails [122]. 21 experts were asked about a phishing email they recently encountered. They were asked questions about what they noticed, what they expected, what felt off, any investigations they conducted, how they decided if the email was legitimate, and what actions they took afterwards. Wash discovered that expert users notice and mentally accumulate discrepancies in an email until they are made suspicious, at which point they conduct an examination of the email and make a conclusion. This process fails when the user does not become suspicious or incorrectly assesses the evidence. Wash contends that those with limited technical expertise are more likely to fail in this process as they may not know the correct characteristics to examine, leading to an inadequate evaluation of the email.

On the other side of expertise, Downs et al. examined security-naive users' sensitivity to features typically used to identify phishing emails [26]. The researchers surveyed 20 individuals, asking them to examine 15 emails. They found that, while participants identified features that commonly characterise phishing, they misunderstood and misused these features when evaluating legitimacy. Further, the participants also focused on unsound features that are not reliable indicators of phishing, such as interpreting

the text of the email body. The researchers concluded that a user being aware of the existence of typical features within phishing emails is insufficient - the user must also understand these features and why they are useful in the evaluation process.

Dhamija et al. assessed 22 participants in their ability to differentiate between genuine and phishing websites [19]. Their findings echoed that of Downs et al., in that even the most careful and knowledgeable users incorrectly evaluated indicators of trust, such as padlock icons within browser search bars, in their estimation of legitimacy. These findings are further corroborated by the more recent work conducted by Kirlappos and Sasse [51], who found that users misinterpret symbols of trust and instead use their own heuristics to evaluate legitimacy which are often unsound.

While each of the above studies had a limited number of participants, Wash et al.'s survey with 297 participants, selected to form a demographically representative sample, corroborates their results and mitigates potential concerns about the generalisability of their conclusions [123]. The participants were surveyed using similar methodology as Wash [122]. Wash et al. found that the majority of participants were less likely to notice 'conclusive distinguishers' - features that can be used to definitively determine an email's legitimacy, such as attachments or URLs. These findings echo that of the studies previously mentioned. Further, Wash et al. found that users utilise unique information that technical filters do not use when determining their estimation of an email's legitimacy. Due to this, they suggest that instead of asserting behavioural changes in users, future work should aim to "strengthen the capabilities that are unique to humans".

Finally, In Harrison et al.'s study involving university students, *all* 113 participants failed to notice the typographical errors in a simulated phishing email [41].

In summary, these studies emphasise that users often misinterpret [19, 26, 51] and overlook [41, 123] information that may help them assess an email's legitimacy. Additionally, they focus on unsound indicators of legitimacy in their evaluations [26, 51].

To build upon this research, a potential study may evaluate the effect of presenting reliable indicators of phishing to users automatically. By highlighting the information users should look for, and assessing the legitimacy of that information for users, the human error involved in the evaluation process may be reduced. It may also direct users away from focusing on features that are unreliable and enhance the unique knowledge users provide. If this is indeed the case, allowing users access to a system that performs this function would reduce their susceptibility to phishing.

To conduct such a study first requires an understanding of the methods previous research has employed to quantify user susceptibility. Through investigating this previous research, the resulting study may take on a more informed approach to quantifying the effect of the aforementioned system.

2.2.3 Methods of Quantifying User Susceptibility

Numerous studies have attempted to quantify user susceptibility and what may influence it. The methods with which these studies accomplish this varies widely across the

literature, including measuring through classroom tests [91], asking participants to classify a list of emails [131], or by sending participants simulated phishing emails to their personal inboxes [99].

2.2.3.1 Embedded Training

Several studies have investigated the concept of 'embedded training'. This training is conducted by sending simulated phishing emails to users' inboxes to determine if they incorrectly interact with them, such as clicking on a URL leading to a website controlled by the researchers.

Using this method, Siadati et al. surveyed 19,180 participants who were each sent 6 simulated phishing emails over an interval of 8 months [99]. The researchers collected data on the the rate at which users clicks on the URL within the simulated phishing email, known as the 'click-through-rate' (CRT). Similarly, Caputo et al. surveyed 1,359 participants, sending each participant 3 simulated phishing emails across 8 months [14].

The advantage of studies with this design are many, including that the materials presented are within a realistic environment that is familiar to the user - within their daily routines and within their actual email inboxes. This is beneficial considering the discussion in Section 2.2.2. However, it is evident that such studies require substantial time and resources. Considering this paper's constraints in resources and time, it is important to instead consider studies that have been conducted with more modest resources.

2.2.3.2 Other Methods

Some studies have explored a classroom approach wherein participants are asked to complete a test. Robila and Raguci involved 48 students in an anti-phishing training course [91], finishing with a phishing IQ test. This test involved 12 emails, 50% phishing, which were shown as images, denying the users from using any interactive analysis they may have conducted on real emails, such as hovering over links to highlight href URLs. The participants were asked to classify each email. Lastdrager et al. similarly employed a classroom approach [60]. They studied 353 children, splitting them into control and intervention groups. The intervention group received a 40-minute anti-phishing presentation. Both groups were asked to complete a test of 10 questions about phishing to measure their susceptibility.

There also exists a more direct approach wherein the participants are asked to classify a range of emails presented to them. Klietman et al. surveyed 150 undergraduate psychology students at the University of Sydney [52]. They developed a 'Phishing Detection Task' wherein 40 emails, 50% phishing, where shown in a randomised order to the participants. These were also shown as images as in Robila and Raguci [91]. Kietman allowed the users to rank each email on a 0 to 100 scale of malice, with with 0 denoting the user is certain the email is not malicious at all to 100 being definitely malicious. The users were also asked to state the confidence of their choice on a six-point scale.

Zheng and Becker employed a similar methodology to Kietman [131]. They sought to investigate whether users effectively utilise email header information in determining

email legitimacy. They recruited a demographically representative sample of 252 participants in the United Kingdom. They exposed the participants to a randomised list of 47 emails in a simulated email inbox where each email was fully rendered in HTML. The participants were asked to process the emails if they were an executive director at a fictive company. Each email was to be given a label from a list of labels, such as 'Needs boss' attention' or 'Discard', specifically created to further avoid the framing effect.

These other methods share similar advantages and flaws. The materials are shown to the user in a single session, offering a time-saving advantage. However, there are also a number of significant limitations: The list of materials are shown outwith a user's routine and may not reflect how their actual behaviour would be. Further, the users must also be informed that their task is to classify the materials into legitimate or phish.

2.2.3.3 Summary

Within these papers, a common practice is observed: A quantitative survey is conducted wherein the participants are divided into two [60, 91] or more [14, 99, 131] groups. The groups are then asked to classify a list of both legitimate and phishing materials - allowing comparisons to be made on the correctness of the groups in labelling the materials. This reveals how phishing susceptibility varies across the experimental conditions tested.

However, it is evident from the subset of studies discussed that there is a lack of consensus on what methodology to employ when measuring susceptibility. Discussing these studies is nonetheless beneficial in informing the potential approaches to a similar study.

2.3 Summary

Users have a crucial role in determining the success of phishing attacks. However, research has demonstrated a number of flaws in the user when they are evaluating an emails legitimacy. Users fall victim to email phishing attacks when they do not perform an active evaluation of the email in question. Further, when they do perform such an investigation, users often misinterpret the information they identify, in addition to including information that is unsound in their evaluation process.

Thus, in the subsequent chapters I will outline a system that addresses the concerns raised by this research through assisting the user in their evaluation process. By presenting additional information to the user, they may be prompted into a more discerning evaluation of the email in question, or it may reduce the likelihood of the user misinterpreting the evidence that an email is phishing by performing the evaluation of said evidence for the user.

Informed by the understanding gained by examining similar studies, I will quantitatively evaluate said system with the aim of determining if this system will make a beneficial addition to phishing defence framework.

Chapter 3

Previous Work and Extensions

In this chapter, I will provide a concise summary of the relevant work I completed in the previous year [95]. I will then briefly discuss the expansions upon this work conducted this year, including additional systems I developed that demonstrate different use-cases for the email analysis module.

3.1 Email Analysis Module

The main body of work conducted in the previous year was that of the email analysis module. This module is capable of parsing an email's technical information and evaluating whether that information is indicative of malice. This information can then be used to evaluate the legitimacy of the email in question.

Motivated by a similar review of the literature as in Section 2.2, an analysis of the technical and social vulnerabilities emails possess was conducted. By investigating technical information and background literature, I discovered a number of ways that emails can be exploited for the purposes of phishing attacks. I summarised these vulnerabilities in Section 2.1.1.

With these vulnerabilities identified, I incrementally implemented ways to detect and evaluate each vulnerability with the information found in EML files. The email analysis module is capable of:

- **Authentication** - The module evaluates the headers containing the 4 authentication protocols to determine the authenticity of the email, and verifies that the protocol headers themselves have not been edited.
- **Sender Analysis** - the module checks the sender's email address against a list of publicly available domains such as *@gmail*, and evaluates whether the sender is outwith the University of Edinburgh network.
- **Attachment Type Analysis** - the module checks all attachments against a known list of potentially malicious file types.

- **Language Analysis** - the module performs basic natural language processing through considering the following features:
  - Phishing Keywords - The module enumerates the occurrences of words within the email that also appear in the list of common phishing words (keywords) compiled by Bergholz et al. [8].
  - Misspellings - The module enumerates the number words within the email that are misspelled.
  - Emotions - The module uses the text2emotion Python module to evaluate the emotions of the email being analysed.
- **URL Analysis** - Given its prevalence in phishing, particular consideration was placed upon creating a sophisticated analysis of URLs. Firstly, the module identifies and enumerates each URL, as well as unique domains within the email. Then, it evaluates the following:
  - Domain Age - The module queries the WhoIs database for the domain's registration date and uses that to determine the domain's age [40, 94].
  - Domain Popularity - The module looks at whether the domain is listed on the security-orientated list of most popular domains created by Pochat et al. [62, 63].
  - Domain Name System BlackLists (DNSBLs) - The module queries 49 DNSBLs to determine if the domain is blacklisted [24].
  - Domain Mismatch - Each URL in an anchor tag href is compared against the anchor tag's text-content to evaluate whether the domains match. Each domain is also compared to the sender's email domain to evaluate whether the domains match.

Each of the above features were used to quantitatively examine two large email corpora: 2,239 verified phishing emails and 4,279 verified legitimate emails. This informed how phishing emails and legitimate emails differ across each feature, identifying whether the presence of certain features - for instance, a mismatched URL - was normal in legitimate emails or indicative of phishing emails.

Further, this corpora analysis informed the creation of a basic logistic regressor. This was used to provide an overall estimation of a given email. Depending on the information obtained, the email is assigned to one of three categories: safe, suspicious, or malicious. However, this logistic regressor was identified to have significant limitations and is not currently fit for a use in a real-world setting. Whether the logistic regressor should remain a part of the email analysis module going forward should be considered. It should be improved through the examination of larger, more up-to-date corpora or entirely replaced by another method to create the overall estimation for an email.

Downs et al. showed that users struggle to understand technical information [26]. Thus, I conducted a brief, iterative design [102] process wherein I created a HTML document that the module dynamically populated with its output information. This allowed the information to be shown in a readable format with explanations of what the information

is and its relevancy in evaluating email legitimacy. Users are then able to read this information and may use it in their evaluation process.

3.2 Extensions Upon the Previous Work

I have completed a range of extensions on the email analysis module created last year. The following sections briefly describe the motivations behind these extensions, their implementations, and limitations.

3.2.1 Additional Features

Throughout the previous year, the email analysis module has been extended to include 2 additional features: telephone number analysis and financial keyword enumeration. Further, the attachment analysis feature has been improved upon.

3.2.1.1 Telephone Number Analysis

Another possible vector of attack within a phishing email is that of a telephone number. Instead of using a URL - an attack that involves setting up a domain and making that domain look legitimate to both filters and users - this attack revolves around convincing the user to call a phone number - a system of communication that possesses its own security flaws that are beyond the scope of this body of work. If the recipient is convinced by the email, they may call the number within where the attacker(s) are then able to continue the social manipulation of the recipient and progress the attack. This vector of attack is on the rise in phishing emails [92, 100]. Further, Carruthers et al. found that email phishing campaigns where the attackers follow-up by calling the recipient succeed more often than they fail [104], demonstrating that this is a powerful tool for social-engineering attacks. Thus, I extended the email analysis module to identify, analyse and evaluate phone numbers it detects within the text of an email through querying the WhoCalled website [126], which aggregates user reports to determine the legitimacy of a phone number and who the number belongs to.

3.2.1.2 Financial Keyword Enumeration

My colleagues at the TULiPS lab have been conducting research into how to categorise phishing attacks into classes, for instance, file sharing scams. One category they identified was financial scams, wherein the attack revolves around impersonating an institution associated with financial services or transactions, such as a bank or tax collecting agency. To that end, I incorporated a list of financial keywords - that is, words relating to money, currency, banking, etc. - into the email analysis module, much in the same way it does phishing keywords. The email analysis module uses this list to determine whether the email is of a financial nature. This could be used to detect financial scams. However, unlike the phishing keywords which were derived from previous research [8], this feature has not been quantitatively verified in any way. Therefore, it is difficult to evaluate its usefulness nor its soundness at this stage. It may be possible to test this feature through evaluating sets of financial vs non-financial

emails and further by financial phishing emails vs non-financial phishing emails. This testing should take place before the feature is used in a security setting.

3.2.1.3 Attachment Analysis

To mitigate the limitations in the attachment analysis implemented last year, the email analysis module now includes an additional list of the file types most commonly used by phishing attacks, but are not inherently dangerous, identified by TrendMicro [109]. These are used to warn the user of potentially malicious files if they are present in the email. It is possible to more rigorously examine attachments through analysing the content of said attachments - a potential avenue of extension to the module.

3.2.2 Extensions

Additional work has been completed on creating tools that use the email analysis module as part of their functionality.

3.2.2.1 Inbox Automation Script

Many organisations employ a dedicated team of experts - known as a Security Operations Centre (SOC) - who help users identify phishing emails by responding to emails they report [27]. However, users over-report emails to such a degree that the false positive rate for reports was 61% in 2021 [21]. This places a significant burden on SOCs, leading to substantial delays in returning a full report to a user [4, 32]. SOCs typically spend 17 to 25 minutes inspecting and responding to potentially malicious emails [32]. However, the time it takes for a phishing attack to be successful is often short, ranging from seconds to minutes after the attack is initiated [115].

To address these shortcomings, I developed an 'auto-responding' script. An administrator can use the script to login to any email address via the command-line interface, a function I implemented through employing the IMAP. The email address thereafter functions as an automated inbox. A user may attach an EML file to an email and send the email to this address. The script will automatically respond to the email using the SMTP with the analysis of the attached EML[1].

This script enables the email analysis module to provide its analysis in response to reported emails within an organisation. Thus, this has the potential to alleviate the burden of email reporting on SOCs by giving the user additional, contextual feedback and advice while the SOC responds to the report. Future work may employ the script and evaluate its effectiveness in augmenting SOCs.

However, while this script has been tested for analysing one email at a time, it has not been tested for handling an inbox with a substantial volume of incoming emails. I suggest the script be tested more thoroughly before employing it in a security setting.

[1]The following address is currently automated using this script to demonstrate this functionality: `mailto:seanstraintesting@outlook.com`. n.b. this uses the response design created last year.

3.2.2.2 Microsoft Outlook Add-In

In the previous year, the creation of an extension (known as an add-in [18]) to the Microsoft Outlook app was attempted. Microsoft Outlook is the email client employed at the University of Edinburgh. This was attempted with the intention of investigating whether it was possible to incorporate the email analysis module with Outlook. This would allow, if completed, any student or staff member at the University to use the email analysis module and incorporate it into the UoE's security framework.

This was attempted again this year with significant progress. As evidenced by Appendix F, the Microsoft Outlook Desktop Application was extended to include a new button with the intended function of allowing the user to report the email to the UoE's SOC. Upon clicking this button, the user is presented with a short survey, allowing them to self-report what made them suspicious of the email they wish to report.

Kirlappos and Sasse found that users respond positively to and place more trust in designs they have seen before and are familiar with [51]. Thus, keeping trust as a central focus, the design of the survey adopts the formatting and colour scheme used by the UoE 'MyEd Student and Staff Portal', a service used by all staff and students at the UoE. The email analysis module's output also adheres to this design. This was intended to create a consistent experience for the user.

However, as in last year's attempt, the Outlook email client was found to obfuscate important information, such as authentication headers within emails. Further, as Outlook is a professional tool, it proved difficult to work with as it places a number of demands on add-ins. The add-in needs to send emails on the behalf of the user, but the process of obtaining the required permissions proved difficult and could not be completed in a reasonable timeframe. As such, the add-in is incomplete. The source code behind this extension has been shared with my colleagues at the TULiPS lab who intend to complete this work at a later date.

3.2.3 Improvements

3.2.3.1 Technical Improvements

Within last year's work, I identified a number of technical limitations within the email analysis module. I researched various software design principles in order to improve upon the system [33, 73, 114]. The module was highly coupled with the output design; without the appropriately formatted template HTML, the analysis would not complete. The module was refactored to allow the information it retrieves to be collected without the need for an output design to be created. This allowed the module to be used elsewhere. Further, the module was refactored to be modular. Certain types of results, such as from language or URL analysis, may be removed from the analysis process should the user desire. A basic logging system was introduced to display the module's results in an intuitive way to the command-line interface. This allows a user to see easily what the module has calculated.

3.2.3.2 Design Improvements

The output design of the module has been improved upon by my colleagues at the TULiPS lab. Through conducting a series of usability workshops, they created and incrementally improved upon a research-informed design supported by user-feedback.

I decided to incorporate their final design into the email analysis module. This design was a static image that required conversion into a HTML document that the module could dynamically edit. Figure 3.1 displays the comparison between the final TULiPS design and the HTML document I created based off of it.

I implemented some changes to the design created by TULiPS. For instance, the TULiPS design asks the user to input some information under the 'Disagree with the Classification?' header, such as whether they have clicked on a URL within the email. It was envisaged that this would allow the design to further contextualise its feedback to include the information the user provides, for instance, it could provide advice on what to do if a malicious link is clicked should the user report that had happened. However, as the HTML document may be delivered through email, no JavaScript may be used in the document [127]. This makes incorporating user inputs difficult. As such, I have removed this feature from the HTML document. This does not take into account the potential use cases where the HTML document may be shown outwith an email - such as if it were to run within an email client - in which case it would be allowed to have JavaScript. As such, it may be possible to have two separate HTML documents, with one including this user feedback. This would require extending the email analysis module to take such feedback into account and is a potential improvement that could be implemented in future.

The TULiPS design envisaged that only the 3 most 'important' heuristics measured would be included in the analysis for a given email. However, determining this information is a difficult task. Currently, the module uses a logistic regressor as described in Section 3.1. The module could use the 3 heuristics that contributed most to the input to the regression function for phishing emails or that contributed least for legitimate emails. Unfortunately, this regressor has significant drawbacks that were noted last year [95]. As such, the HTML document shows all information the analysis discovers and the most important feature is chosen manually for each analysis. This process will need to be improved in future, be it by improving/replacing the logistic regressor, or otherwise.

3.3 Summary

Within this chapter, I have described a system that automatically analyses features of emails that are effective in identifying phishing emails [95].

To build upon the previous work in phishing literature and my work last year, I propose a study that will quantify the change in phishing susceptibility in users when presented with the information from this system. I hypothesise that this system will allow users to become more engaged in the evaluation process and direct them towards more reliable indicators of legitimacy. Thus, the user will be less susceptible to phishing attacks.

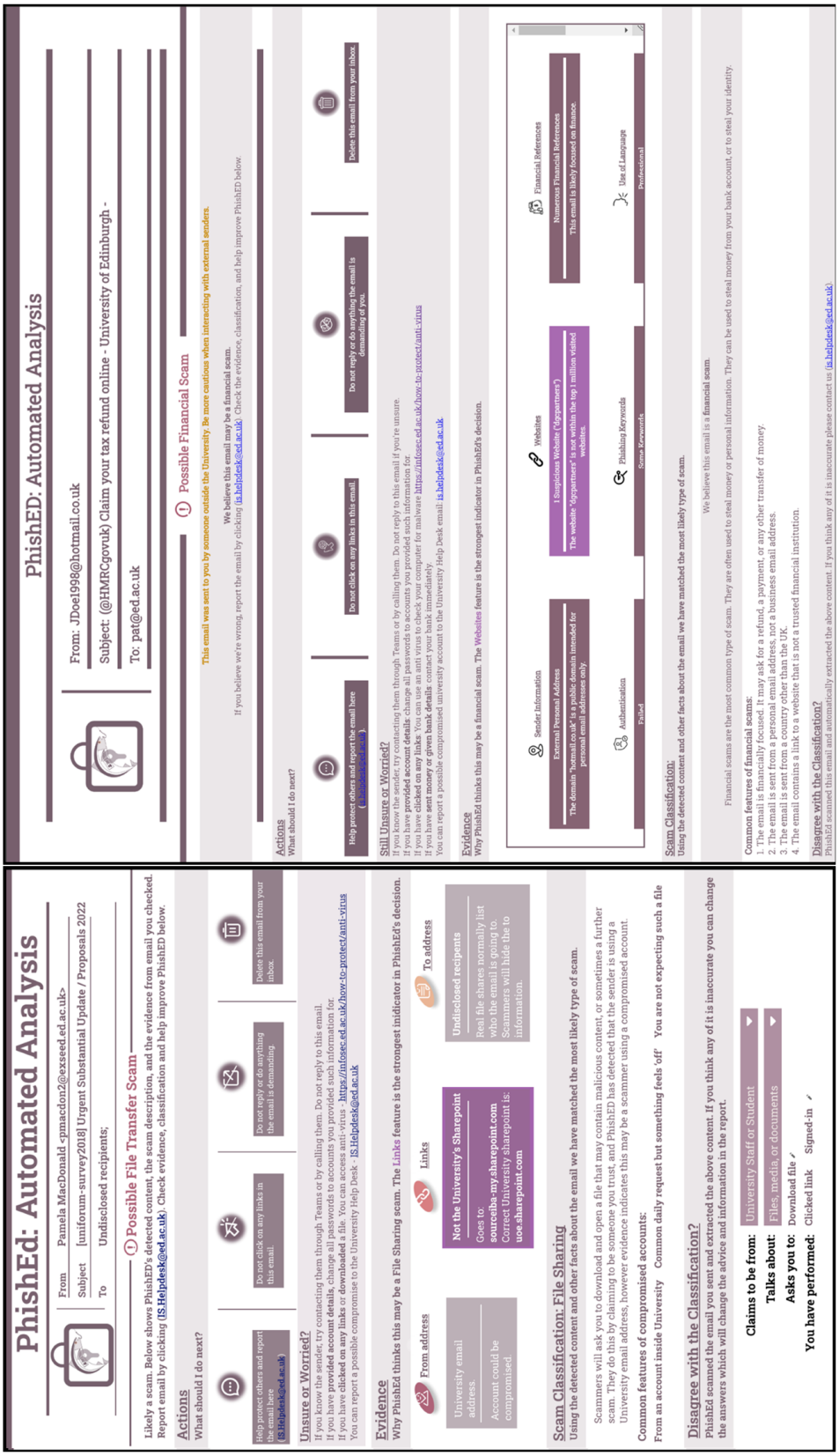


Figure 3.1: An example of the final design created by TULiPS (A) and corresponding HTML output I created based off of this design (B). Rotated 90° anticlockwise.

Chapter 4

Methodology

I hypothesise that users with access to the analysis provided by the email analysis module will have greater phishing email detection ability than those without. In this Chapter, I outline a study that aims to evaluate if this is indeed the case. I will describe the experimental conditions, material selection, and the approach taken with the analysis of the results.

It should be noted that the proposed study - while a complete study in its own right - also serves as a pilot study for a larger scale study to be conducted by the TULiPS lab later this year. Pilot studies have been used previously in phishing literature as a way to test systems and highlight concerns, but also as a way to receive preliminary data to inform the larger study [98]. Thus, as a pilot study - in addition to seeking an answer to the hypothesis - this study also aims to form the groundwork for the larger study by testing the systems used, identifying potential flaws, and solving highlighted issues to make the larger study more effective.

4.1 Overview

As discussed in Section 2.2.3, a study that follows the style of embedded email studies would possess a number of advantages. Such a study would be achieved by incorporating the email analysis module into the participants' email clients. The materials would be presented within a realistic environment that is familiar to the user - within their daily routines and within their actual email inboxes. Further, this study design would allow for a number of useful statistics to be collected, such as how often the tool was used in a user's usual routine. However, such studies are are conducted over a significant timeframe, such as over many months [14, 99]. Further, it would require the tool be incorporated into the users' email clients. Given the aforementioned difficulty in achieving this with Microsoft Outlook, discussed in Section 3.2.2.2, this precludes the embedded method until these issues are solved or an alternative is proposed.

Instead, I was supplied access through the TULiPS lab to the source code used in Zheng and Becker's study [131], with their consent. This code provides the implementation of a website that simulates the experience of processing emails in an inbox. Users are

then able to classify/label each email (for instance, as phishing) in a list of emails in this simulated inbox, the website then records the results. I identified this website could be modified to evaluate the email analysis module by incorporating it into the website.

This method has a number of advantages: It would allow a list of materials to be shown to the user remotely, data may be collected both automatically and quickly, once set up there is no further intervention required, etc. . However, there are also a number of significant limitations: The list of materials is shown outwith a user's routine and may not reflect how their actual usage of the tool would be, the website must be compatible with a wide array of browsers as opposed to just one email client, etc. . Overall, I decided using the website and the methodology it entailed was preferable due to the advantage it provided in time; I envisaged it would take less time to modify and finalise the website for the needs of the study than create the aforementioned embedded study. That being said, given the advantages of an embedded study and the limitations of this method, I suggest an embedded study may be the preferable choice of study design for future work.

The website provided by Zheng and Becker has been modified for the purposes of the methodology described below. This is discussed in the subsequent Chapter 5. Images of both the original website and modified website are supplied in Appendix B. The modified study website is accessible through the UoE's School of Informatics local network or the UoE's Virtual Private Network at the following URL: `http://129.215.10.154:8080/`.

4.2 Study Design

This study closely mirrors the design and methodology of Zheng and Becker [131]; all participants will be asked to classify a list of emails in a simulated inbox environment. The following subsections detail the task flow, the experimental conditions, and, lastly, the email selection process and resulting email set used in the study.

4.2.1 Task Flow

The website will guide each participant through a four-stage process: 1) Consent - They must give their consent to to complete the study. 2) Instruct - They are shown an instructions page informing them of their task and how to correctly use the simulated inbox. 3) Task - The participants will all be asked to label every email in the simulated inbox. 4) Survey - Once this labelling process is complete, the participant is asked to complete a 6-question survey.

During the task phase all participants have access to, for each email, the *Subject*, *To*, *Date*, and *From* headers, as well as the full email body rendered in HTML. All URLs were altered to redirect to a blank page. The order of emails in the inbox is randomised for each participant to control for the order-effect bias [86]. Participants must label each email as either 'legitimate' or 'phishing'. Further, following a simplified approach to Klietman et al. in measuring user confidence [52], participants must state their confidence in their decision using a two-point scale: 'confident' or 'cautious'.

4.2.2 Experimental Conditions

Participants will be randomly assigned between two groups: 'Control' and 'Experimental'. The consent phase is equal for both groups. In the instruct phase, the groups are given different information. As the system under evaluation is novel, participants in the Experimental group must be informed of its usage and purpose. In doing so, given the discussion in Section 2.2.1, the group is primed into detecting phishing. To maintain the comparability between the two groups, the Control group must also be similarly primed. Therefore, the Control group is given access to preventative anti-phishing advice[1].

The experimental group will have access to the email analysis module's precomputed analysis of every email during the task phase.

After completing the task phase, all participants must complete a 6-question survey. This survey can be seen in Appendix D. Question 1-5 are mandatory questions. Question 1 asks for the participant to non-anonymously submit their output file[2] to the cloud storage supplied by Microsoft Forms. Their name was immediately removed to re-anonymise the response. The following 4 mandatory questions were demographic questions. The participants were asked to supply their age bracket, gender identity, current level of study, and nationality. The last question was an optional additional feedback question.

4.2.3 Email Set and Selection Methodology

Zheng and Becker used 47 emails in their study. These emails were selected to match their persona of 'Alex Carter' - an executive director at a fictive company. Personas are a concept in the field of Human-Computer Interaction (HCI) [70]. According to Martin et al. [70], attempting to create a design that works for everyone often leads to "unfocused or incoherent solutions". The email set is susceptible to such concerns. Each email inbox is unique, with a unique list of emails coming from a unique list of sources depending on the user and the communications they receive. Thus, attempting to create an inbox to suit every user will result in an unfocused solution. This problem is mitigated by personas [70], which allow the consolidation of design goals to suit an archetypal, yet fictional, user; in Zheng and Becker's case, this was Alex Carter.

The same email list and persona that Zheng and Becker used may have been used for the purposes of this study. However, for the abundance of caution in keeping within the ethical agreement with the University of Edinburgh, my colleagues at TULiPS decided the list of emails should be wholly changed to emails collected by TULiPS and I, along with the persona those emails were prescribed to fit.

Determining the persona requires identifying the archetypal user. As the study would be introduced to students at the University of Edinburgh, it would follow that the persona should emulate a typical user in this group. Thus, I created the persona of Pat Smith - a 5th year undergraduate master's student at the University of Edinburgh. The inbox would be shaped around usual emails a 5th year student would expect to receive, such

[1] Specifically, the advice given was extracted from the University of Edinburgh at the following URL: `https://infosec.ed.ac.uk/how-to-protect/think-before-click/about-phishing`.

[2] Microsoft Forms prohibits anonymous file uploads.

as job offers, university communications, or emails from accounts Pat may be registered with, etc. . This persona allows a much more focused list of emails that the participants can relate to.

The number of emails used in the study was decided by a consensus agreement with my colleagues in the TULiPS lab and I, with advice from Zheng. There is arguably no one single email that can alone determine a user's susceptibility to phishing; Klietman et al. recommended that users be tested across a range of materials [52]. Thus, it was decided that the study would contain 16 emails. This amount, reduced from the 47 of Zheng and Becker, would allow participants to complete the study in a shorter timeframe, a factor considered advantageous given the voluntary nature of the study. However, this limited number reduces the range of phishing emails that can be measured, reduces the likelihood of detecting significant effects [6], and is less than the number used in many similar studies [52, 98, 131]. The future study should seek to include more emails in its set to match existing literature and to improve the study's statistical significance.

Of these 16 emails, 5 were phishing. This proportion ($\approx 31\%$) was informed by Zheng and Becker, who described their apportionment of 17% phishing as a 'more naturalistic phishing proportion' than the 50% used in other studies [19, 52, 91]. However, they do not provide any justification for this claim. Furthermore, a 17% proportion would yield only 2-3 phishing emails within the set of 16, which was considered insufficient by my colleagues in the TULiPS lab. Thus, 5 phishing emails were chosen as a compromise between the proposed 'natural proportion' and an acceptable number of phishing emails for testing purposes. A future study may look to use a proportion closer to that of Zheng and Becker or identify a more justified proportion based on real-world phishing email prevalence or future research findings.

The 5 phishing emails were chosen by a colleague at the TULiPS lab. They used the PhishScale methodology, devised by Steves et al. [105], to select from the phishing emails collected by the TULiPS lab. This resulted in a set of phishing emails that included a range of phishing cues and that suit the persona. I selected the 11 legitimate emails from my personal inboxes. I aimed to create a set of emails that originated from a range of sources that suited this persona. Thus, I chose university emails, emails relating to personal accounts, and generic emails. This process was not informed by any particular, research-informed methodology. Thus, the selection of legitimate emails should be improved on in future by using a more informed approach.

Each email is given a unique index. The phishing emails are indices 1, 2, 5, 6, and 10.

These materials were evaluated using the email analysis module. The module correctly evaluated all but one email, email 14, which the email analysis module incorrectly determined to be a phishing email.

Images of the email body and corresponding analyses of emails that were identified as notably interesting given the study's results (Chapter 6) are supplied in Appendix E. Images of all other emails are supplied in the supplementary materials submitted with this report.

4.3 Analyses of Results

The study was guided by the following hypotheses [64]: **1a)** Participants in the Experimental group will correctly identify more phishing emails compared to the Control group. **1b)** Participants in the Experimental group will correctly identify more legitimate emails compared to the Control group. **2a)** Participants in the Experimental group will be more confident in their decisions when labelling phishing emails compared to the Control group. **2b)** Participants in the Experimental group will be more confident in their decisions when labelling legitimate emails compared to the Control group. **3)** Participants in the Experimental group will take longer to process all the emails in the simulated inbox.

To test Hypotheses 1a, 1b, 2a, and 2b, Fisher's exact test was employed [71]. Fisher's exact test is a type of statistical test that allows for the measurement of the relationship between two variables. Other tests of this type exist, such as the χ^2-test that Zheng and Becker employed. However, I decided to use Fisher's Exact test as this study collected a much smaller sample size (N=22) than Zheng and Becker (N=252). Fisher's Exact test is more suited to smaller sample sizes, as stated by McDonald [71]. In the future study, this methodology may be revised if the sample size is increased.

The Fisher's exact test was employed with equal expected detection rates and confidence for the two groups under the null hypothesis. The standard value of significance ($p = 0.05$) was used. The detection rate for each group is the number of participants who correctly labelled an email divided by those who incorrectly labelled. A 2x2 contingency table with columns 'correct' and 'incorrect', and rows 'Control' and 'Experimental' will be created, allowing for the Fisher's exact test to be conducted. Confidence is determined with an equivalent process using 'confident' and 'cautious' labels. As the participant may supply a list of labels for each email, the final label assigned to each email is considered the participant's decision.

Emulating Zheng and Becker [131], the precision and recall of each participant will be calculated for both phishing and legitimate email sets. For each set, the precision is determined by the proportion of correctly labelled emails in that set out of all emails the participant labelled as belonging to that set. The recall is the detection rate for each participant across the email set, and is used to determine the SD of the detection rate for each group.

For Hypothesis 3, the time to label each email for each group will be compared, alongside the time to label the entire inbox.

In addition to evaluating the hypotheses, I will perform an exploratory analysis [110] of the usage patterns of the email analysis module within the Experimental group. Specifically, the frequency of module usage across all emails, as well as between phishing and legitimate emails, will be analysed. I will measure average use through both the arithmetic mean and median. Both methods are suited to different distributions of data, and the distribution cannot be determined in advance. I will measure the dispersion through the standard deviation (SD). Further, I measure the frequency users label an email the same as the analysis module classifies the email to identify if users comply with the module, even if it is incorrect.

Chapter 5

Implementation

This chapter outlines the process of modifying the website provided by Zheng and Becker website for the purposes of the study proposed in Chapter 4. Firstly, I performed an investigation into the technical details of the original website to inform the modification process. Then, I incorporated the email analysis module and altered the relevant functions of the website accordingly. Finally, I conducted a series of tests to ensure the modifications had not harmed the functionality or usability of the website.

5.1 Modifying the Survey Website

5.1.1 Overview

The source code provided by Zheng and Becker contains a series of webpages that utilise the open-source Vue JavaScript framework [120]. These webpages guide a participant through the following phases: The participant consents to their data being collected (consent), is given instructions on how to use the inbox (instruct), and is then shown a list of emails to label (task). The source code also contains the code necessary to process the labels each participant assigns to each email, and deliver the results per participant to a database. These webpages are shown in Appendix B.

During the task webpage, the user processes a list of emails. Each email has been separated into two parts: the header and body. The majority of the header information is removed, only the *Subject*, *From*, *To* and *Date* headers are preserved. Each email's preserved headers are presented in a column on the left of the webpage, as in the Microsoft Outlook email client. The user is able to click on each of these entries which reveals the corresponding email's body content on the right of the webpage.

The simulated inbox uses a list of inline frames (iframes) [75], a HTML element that loads another HTML document within the document, to display the bodies of the emails. This is beneficial as many email bodies are already a HTML document, and allows for additional user interaction with the email, such as closer inspection of URLs which is impossible using static images. The inbox has of a list of these iframes - one for each emails' body. These iframes are hidden by default and are only displayed to the user

once they have clicked on the corresponding entry with the preserved headers on the left.

The Vue script employs a meta-data file of the file type JSON. This file contains for each email: The preserved header information, the filename of the corresponding email's body as a HTML document, as well as other information to be used later, such as a unique ID per email and whether or not the email is a phishing email (known as its 'truth label'). This file is used to dynamically create the inbox, with each entry on the left column being associated with an element in the meta-data file.

5.1.2 Usability

Usability is an important consideration when developing software. Nielsen, a prominent authority in the field of HCI [36], notes that while 'utility' is about the features a software provides, 'usability' is about how easy those features are to use [37]. Nielsen stresses that people will not interact with a website that does not have an acceptable level of usability. Thus, any changes to the website should be evaluated from a usability perspective to ensure the change does not render the completion of the user's activity unnecessarily difficult. Nielsen established 10 principles that can be applied to improve the usability of websites [35, 79].

Further, I followed the Web Content Accessibility Guidelines (WCAG), which defines how online resources can be as accessible as possible to those with disabilities [128].

5.1.3 Incorporating Email Analysis

The email analysis module extracts a range of technical information, as described in Section 3.1. However, this information must be shown in a readable way for layman users [26]. As such, the module employs a single page output design in HTML which it creates for every email, as described in Section 3.2.3.2. The use of HTML for this output means it can be placed into the webpage through iframes in the same way the email bodies are. Thus, I extended the webpage to include a new column of iframes - one per email - where the corresponding analysis would be shown.

This means the webpage would now be divided into thirds. The leftmost third of the screen shows the email headers, the middle third of the screen shows the currently selected email's body, and the right-most third of the screen shows the currently selected email's analysis.

This immediately raised a usability issue. The webpage was crowded with information and the two iframes of the email body and the analysis were being fit into too tight a space. This caused various overflow issues where the design of the analysis was warped and stretched in unexpected ways to fit the information into the small space.

I considered a range of solutions to this crowding. One solution would be to allow the user to choose to see either the analysis or the email body, not both. This would give each iframe enough space to be shown properly. However, this would force the user to go back and forth between the two iframes as they processed the list of emails. The user must remember the information stored in one iframe as they read the other one.

The analysis heavily relates to the content of the email and is there to assist the user in understanding that content. This goal is made more difficult if they have to remember what that content was. This violates the usability principle of 'Recognition over Recall', the user should not have to recall the email's content as they read the analysis of it. Thus, I rejected this approach.

To solve this, I took inspiration from the LaTeX editor 'Overleaf' [84]. This platform shares a similar issue: The raw LaTeX and compiled document must occupy a limited amount of screen space. LaTeX uses a resize tool that gives the user the flexibility to choose the size of the elements on screen. Their solution conforms to the usability principle of 'Flexibility of Use': It gives the user the freedom to shape the webpage how they choose. Thus, I decided to incorporate this functionality into the webpage.

This involved adding a fourth, smaller column to the webpage which would be placed between the email body and analysis. It consisted of a vertical bar. The user may click on this bar and drag their mouse to move the bar which would cause the iframes to resize accordingly. I used JavaScript event listeners to complete this. These would keep track of the user's inputs and mouse location when they click, move, and release the mouse. However, this solution ran into a fundamental issue: Event listeners only operate upon the document they have been set to and do not operate within iframes. That means that if the user were to move the mouse too fast and it entered an iframe, the event listener would stop working and this resulted in unexpected behaviour, such as the resizing continuing after the mouse click had stopped.

Instead, I found that it was possible to use the Cascading Style Sheets (CSS) property of 'resize'. By adding this property to the email body iframe, it was possible to allow the user to resize the iframe as they wished. This was a much simpler solution that succeeded without the problems discussed with the Overleaf style resize bar. I decided to revert the changes to include the Overleaf style resizing in favour of this approach. However, this solution had limitations that were revealed in Section 5.2.1.

With this solution implemented, I added a new button to the webpage that would open the analysis iframe on the right when clicked. In addition, I extended the meta-data file to include the filename of the email analysis, allowing each email to have an associated analysis. Thus, the email analysis was incorporated into the website.

5.1.4 Changing the Email List

It is possible to dynamically change the list of emails shown on the website if the list of emails is presented in the correct format. The overview above detailed the meta-data file, preserved headers, and extracted email bodies. This is clearly more complex than a directory of EML files downloaded from an email inbox. How Zheng and Becker created this meta-data file is not described, nor is how they extracted the body of an email and converted it into the HTML documents necessary to display the email. In order to change the emails in the inbox it is first essential to be able to create this information from a given list of emails.

As such, I developed a Python script that - when given a directory of non-phishing emails and another of phishing emails - converts those directories into one directory in

the form required by the website. Given the email analysis module already contains a component that parses an EML file, extracts the headers, and decodes the body information, the new script was able to re-purpose that component. The new script places the relevant headers into the meta-data file and the body into a HTML file. If the body is not already HTML formatted, i.e., is plain-text, it simply wraps the content in a *p* tag and creates HTML a document containing only that tag. The result is a directory of HTML files and a meta-data file with one entry per email.

I encountered a problem when determining the 'height' entry for each email in the meta-data file. The website determines how how much vertical space the iframe of an email should occupy on the page based off of the meta-data's height entry, measured in pixels. This is important, as too small a height would hide information in the email behind an unnecessary scroll-bar, or fill the page with unnecessary white-space, both of which harm usability. However, there is no way to determine the height of a given HTML document without it being computed and rendered first by a browser. Through my communication with Zheng, I discovered Zheng and Becker encountered this problem as well. They completed the process of computing heights manually; this would involve opening each HTML file in a browser and manually inputting the value of the computed height into the meta-data file. However, I discovered this process could be automated through the use of Selenium [96], a web browser automation tool. The primary use case for Selenium is for web browser automation - usually for the purposes of web-scraping or testing. I discovered it could be used to solve this problem. With this module, I employ the following logic: open the document in the Firefox browser, append a short JavaScript script to the document which extracts the computed height, return this value to the Python process, and use the value to fill the meta-data entry of 'height' for that email. Given each browser renders HTML differently, it may be that the height computed by Firefox is not the same as it would be in, for instance, Chrome or Safari. This is a known limitation. It could be solved by including more entries in the meta-data file, one for each browser, and using the relevant entry based on the user's browser. However, given this potential issue was not raised in testing in Section 5.2, I deemed the current process sufficient for the purposes of this study. This limitation must be considered in future.

With this problem solved, the script was capable of automatically creating the meta-data file and HTML documents required to place a list of emails into the website.

Zheng and Becker made the URLs in the phishing emails safe through the moving the URL to the 'title' attribute in HTML. When clicked, this takes the user to a blank page. However, this makes the URL more prominent when hovered, which does not reflect how URL hovering works normally. Thus, I altered this methodology. Instead, I removed the Hypertext Transfer Protocol (Secure) part from the URL. This will still redirect the user to a blank page within the website. This means participants remain able to evaluate the URL by hovering over it, and it more closely reflects actual usage.

I added to this script to include an analysis of each email in the directories, generate the outputs accordingly, and add the output filename to the meta-data entry for each email. This addition meant that the script was able to totally automate the process of setting up the required information for the webpage: meta-data, email bodies, and email analyses.

5.1.5 Changes to User Flow

The prospective study requires that the participants be split evenly into control and experimental groups, as described in Section 4.2.2. These two groups are required to be exposed to two different sets of materials. This in turn requires the website to determine what group a participant is assigned to and change the information shown accordingly. This involves changing the 'user flow'- that is, the set of steps between a participant entering the website and reaching a successful outcome [82] - depending on the participant's group. This functionality was not present in the original source code.

I implemented a random sorter that assigned each visiting participant to each of the groups with equal probability. To verify it was working correctly, I had the randomise function assign 10,000 simulated participants, and found it split the participants evenly within an acceptable margin of error.

The instruct page was expanded to allow for two groups. Both groups are given the same information on how to use the inbox and label emails. The task page was edited to only show the email analysis button if the participant was in the experimental group. This prevented control participants from opening the analysis iframes.

5.1.6 Changes in Data Retrieval

As mentioned in Section 5.1.1, Zheng and Becker supplied the code necessary to send the users' results to a database. Specifically, they made use of the Google Firebase database. I was precluded from using this software due to the ethical agreement I was bound by, which asserted that all user data must be stored on the University of Edinburgh's local network. Thus, an alternative solution was required.

I devised a solution wherein the user's data would be written to a file. I implemented this by writing the information to a file, attaching said file to an anchor tag at a fixed location on the webpage, and initiating a mouse event of 'click' on the location of the anchor tag. This downloads the file directly to the user's device's filespace. I tested this functionality on Mozilla Firefox and Google Chrome on both Windows 10 and Ubuntu 22.04.2, and verified it worked on each. However, this functionality was found not to work on some browsers in Section 5.2. A solution to this limitation would be to use a University of Edinburgh server to host a database. The server could listen to POST requests sent from the website with the user's data stored within and update the database accordingly [77]. In future, I recommend adopting this database approach, providing a universally compatible solution.

As the form of data retrieval was changed to a file downloaded to the participant's device, this necessitated a way for the participant to submit that file. I decided to make use of a Microsoft Forms - an online survey platform [81] - as this software was within the ethical agreement. This approach would also allow further questions to be asked about the participants, described in Section 4.2.2.

I also changed the labels the user could give to each email. Zheng and Becker used the labels of 'keep', 'delete', 'forward', 'archive'. For the prospective study, I simplified this to 'Label as Legitimate' and 'Label as Phishing'. This simplification was implemented

as I predicted that the study would receive few recipients, meaning that too many labels would dilute the data too greatly between the different labels. This would make creating conclusions based off of the data retrieved more difficult.

I changed the labelling process to be stored as a list of decisions for each email, as opposed to only allowing the user to make one, irreversible choice as Zheng and Becker had implemented. This new process allows for a thought process to be stored for each email, for instance, a user may label an email 'legitimate' and subsequently change their mind to 'phishing'. Further, this change implements WCAG Success Criterion 3.3.6. With this change, this thought process would be reflected in the data as opposed to it being deleted as in Zheng and Becker. Further, any time the user clicks to analyse the email, the label list for that email is appended with the 'analyse' label, allowing the data to reflect users' usage of the tool. Finally, I appended to each label the time taken from the user starting the task to creating the label. This would allow for comparisons to be made regarding how long users take to perform certain actions, such as how long it took to label an email.

5.1.7 Summary

Overall, I implemented a series of modifications to the website to allow it to be used for the purposes of the prospective study. I devised and created improvements to Zheng and Becker's source code, including automating some processes and adding to the data retrieved. I encountered several issues in this modification process, and created solutions to each.

The webpages of both Zheng and Becker's original website and the completed, modified website can be seen in Appendix B.

5.2 Testing

With the modifications complete, it was important to verify that these changes did not harm the website's usability and that participants were able to submit their responses without issue. Thus, I began testing the website.

5.2.1 Think-Aloud Testing

Think-Aloud Testing is a method of evaluative, observational testing [70]. It involves a user being given a specific task to complete using a system while they verbalise their thoughts, which are recorded throughout the process. I chose this type of test as it provides qualitative feedback upon the aspects of a system that most regularly frustrates or confuses users [70]. By identifying these frustrations, I could alter the aspects of the systems that cause them and thus improve upon the website.

I gathered 2 students with experience in user-interface design from the University of Edinburgh. Both were 5th year undergraduate Master of Informatics students the University of Edinburgh and had previously created or worked on other websites. I

instructed them to complete the task of labelling all the emails as I noted their thoughts. Their responses are noted in Appendix A.1.

Both participants noted issues with the resizing tool described in Section 5.1.3. It was seen as too small and non-obvious - one participant had to be told the tool was there. Further, the lack of feedback received as the participants labelled emails was a common concern; they recommended some sort of visual confirmation that they had labelled an email. Finally, the issue of the buttons to label emails being placed at the top of the webpage was also noted; this required the participants to scroll to the top of the page each time they wanted to label an email.

To address the first concern, I added a linear gradient at a 45° angle at the bottom right of the email body iframe, where the resize tool button is. This makes the tool clearer[1]. For the second concern, I added a 'background-color' CSS property to labelled emails - light red and light green for labelled phishing and labelled legitimate, respectively. Also, the email list entry for that email is shown with the same image shown in the button used to label it, for instance, a shield with a tick is shown next to emails labelled legitimate.

For the third concern, I implemented the 'sticky header' design pattern [61]. This design pattern suggests a persistent header that has a fixed position on the page, regardless of user-scrolling. This was completed by adding the CSS property of 'position: fixed' to the HTML element.

If I had access to more resources, I would have performed this testing with more participants; 2 participants is a small sample size. Johnstone et al. note that 5-10 participants should be used for think-aloud tests at minimum [49]. Further, as changes were made due to feedback, it may have been prudent to repeat testing to see if any additional issues were noted. Thus, it is likely usability problems were missed. However - even with the limited sample size of this test - actionable feedback was returned.

5.2.2 Web Browser Compatibility

Given the study will be distributed remotely to participants in the form of a URL to the study website, there is no restriction on the browser each participant may use. This necessitates some form of cross-browser testing to ensure no users encounter issues due to their chosen browser. Before conducting this testing, I used the statistical website StatCounter to identify the market share of each browser as of February 2023 [103]. This allowed me to better inform my approach to this testing by identifying the browsers to prioritise to reach the most potential participants.

One way that browsers differ is in the way they render HTML and CSS. Each browser supports a different subset of the available properties of these technologies [13]. To test that the layout of the website does not change in such a way to harm its usability, I performed a screenshot test using the commercially available BitBar tool [11]. This allowed me to verify that the website would load as expected and that the website's layout would not substantially change on a browser. The results of this test can be

[1] This can be seen more visually in Section B.

shown in Appendix A.2. This revealed that the website had minimal layout changes on the most prevalent browsers currently used. The website did not load correctly on Internet Explorer (IE) with an unspecified error being displayed. However, as IE only has a 0.25% market share, I did not investigate this matter further.

While this test verified the website appeared the same across browsers, the test could not verify that the functionality of the website would not change across browsers. For instance, it could not verify all of the buttons always worked the way they should. I performed this testing manually by opening different browsers and attempting to complete the labelling of all emails. I was able to test the browsers of Google Chrome, Mozilla Firefox, and Microsoft Edge, which together constitute $\approx 73\%$ of users' primary browsers. I was able to perform the study to completion on all 3 of these browsers. However, I did not have access to all browsers on all systems. Most pertinently, I was not able to test the second largest browser by market share ($\approx$18.8%): Safari. Safari is no longer available for Windows or Linux [20].

Instead, I asked a colleague at the TULiPS lab to test Safari on their device. They identified that the file download workaround I created in Section 5.1.6 did not work. As I did not have personal access to the Safari browser I could not test the cause of this issue and therefore the issue was difficult to solve. After unsuccessfully attempting an alternative method of issuing the click command to the anchor tag referenced in Section 5.1.6, the limitation of time forced me to move on. Thus, I implemented a method to detect the user's current browser. I used this to add a warning to the first webpage the user sees to inform them not to use Safari. This is not a solution, this limitation should be investigated further before running the larger study given Safari's large market share.

5.2.3 Software Testing

While testing usability is important, so too is ensuring the implemented website returns the correct results. This depends on the script running on the website that captures the user's labels. If this script were to be incorrect, it has the potential to render the results of the study null and void. Bessey et al. note that, if a system is big enough, there will be mistakes in the code [10]. Further, Yasar writes that software testing is 'imperative' in finding these mistakes [129]. Thus, to verify the script was correct, I completed the task phase while manually noting how I labelled each email. I compared this to the output file and verified it stored the correct labels. Further, I timed the length of time it took to complete the task using a digital stopwatch on a mobile device, and verified the time stored in the output file was reasonably close to the digital stopwatch. Finally, I implemented unit tests. These tests ensure that: 1) the time stored was strictly increasing for each label added to the output file, 2) the correct label is added to the file, and 3) labels are not overwritten or removed.

With all functions tested and verified, I considered testing complete. However, I suggest that this testing should be more rigorous before the full-scale study. This testing could involve a larger think-aloud test or additional/alternative usability tests.

Chapter 6

User Study Results

This chapter provides an overview of the key findings and results obtained from the study. The study aims to reach a conclusion upon the hypotheses noted in Chapter 4, and perform an exploratory analysis on participants' usage of the email analyses provided to them. The data collected from the participants is analysed and discussed. Emails specifically noted as interesting are discussed in further detail, and are supplied in Appendix E. All values will be rounded to 2 significant figures.

While analysing the participants' labels, I identified participant 22 had 6 correct labels, which was notably low compared to the median correct labels of 14; the participant with the next worst correct labels had 11 correct labels. To determine if this was statistically probable, I employed Grubbs' outlier test [39]. I chose Grubbs' test over other methods of outlier detection as it is specifically designed for detecting a single outlier in a dataset, given that dataset follows the normal distribution. Given the sample size (N = 22) and the standard significance level (p = 0.05), the calculated value for participant 22 exceeded the critical value of Grubbs' test, suggesting that participant 22's performance significantly deviated from the rest of the sample. Moreover, the Shapiro-Wilk test [97] revealed that the distribution of the participants' correct answers without this outlier is indeed approximately normal ($p > 0.05$), justifying the use of Grubbs' test. Consequently, I have excluded participant 22's results from subsequent analyses. Participant 22 was a member of the control group, resulting in a 10 to 11 split between Control and Experimental, respectively.

Given 21 remaining participants and 16 emails, 336 final labels were recorded, 160 from the Control group and 176 from the Experimental. While the labelling system allowed participants to change the label they assign to an email, i.e., change their mind, there were only 4 occurrences of this, with 2 occurrences per group.

6.1 Exploratory Analysis of Module Usage

The analysis module was used 56 times out of the 176 times an email was viewed by the Experimental group. When the module was used, it was used once per email per participant, with the exception of participant 12, who clicked to analyse emails 13 and

15 three times and two times, respectively. Thus, there were 53 unique uses of the module, resulting in a usage rate of $\approx 31\%$. No participant analysed every email.

4 of the 11 Experimental participants did not use the module on any email. This was despite the users being prompted to use the module in the instruction phase. All 4 of these participants mislabelled at least one email.

Across the Experimental group, the mean number of times the email analysis module was used was 4.82, with a SD of 5.42, and a median of 5 times across the 16 emails. When excluding participants who did not use the module at all, the mean usage increased to 7.57 with a SD of 4.96, while the median remained unchanged at 5. These large SDs suggest that there is a wide range of usage patterns among the Experimental group; some participants used the module quite frequently, while others were more selective.

Between phishing and legitimate emails, usage remained similar. The mean number of times a phishing emails was analysed was 3.40, SD 1.14, median 3. The mean for legitimate emails was 3.60, SD 1.43, median 4. These findings suggest that participants used the email analysis module at a similar rate for both phishing and legitimate emails.

Of the 53 times an email was analysed, the participant agreed with the overall classification of the analysis module 49 times, $\approx 92\%$ of the time. As evidenced by Figure 6.2, 5 of the 11 Experimental group participants mislabelled email 14, the one email it incorrectly classified. None of the participants in the Control group mislabelled this email. I employed Fisher's exact test, which found this difference to be statistically significant ($p = 0.035$). Further, when filtering the experimental group to include only those who analysed the email (4 participants), all followed the classification and mislabelled the email. When considering these 4 participants against the Control group, p is < 0.001.

6.2 User Performance

The labels assigned to phishing emails by the participants are displayed in Figure 6.1. The phishing emails have the indices 1, 2, 5, 6, and 10. The labels assigned to legitimate emails by the participants are displayed in Figure 6.2. Of the 16 emails, 9 were mislabelled at least once.

6.2.1 Phishing Emails

6.2.1.1 User Performance

According to Hypothesis 1a, it would be expected that the detection rate of the Experimental group for the phishing emails would be higher when compared to the Control group. The control group were correct 88% of the time (SD = 0.17), compared to the Experimental Group at 91% (SD = 0.14). This difference is not statistically significant (odds ratio = 0.73, p = 0.75). As such, there is currently insufficient evidence to confirm Hypothesis 1a.

Emails 5 and 6 were the most often mislabelled. For email 5, 3 of the 10 participants in the Control group mislabelled this email, as opposed to 1 of the 11 Experimental participants. However, the amount of data collected is insufficient to determine if this is

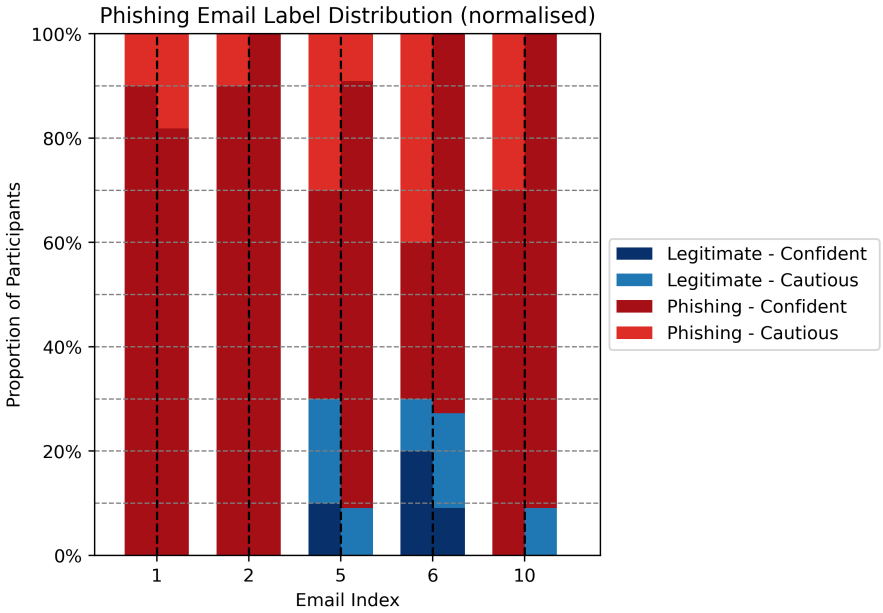


Figure 6.1: Labels assigned to each phishing email in the email set by the participants. Each email has two bars: The left bar indicates the labels assigned by the control group, the right indicates the labels assigned by the experimental group.

statistically significant (odds ratio = 0.23, p = 0.31). An equal number of participants mislabelled Email 6 across both groups.

The precision for phishing emails of the Control and Experimental Group was 0.76 (SD = 0.17) and 0.94 (SD = 0.08), respectively. This suggests that when a participant in the Experimental group thought an email was phishing, they were more likely to be correct than the control group, and this high level of precision is consistent across the group.

6.2.1.2 Confidence

According to Hypothesis 2a, it would be expected that the confidence of the Experimental group for the phishing emails would be higher when compared to the Control group.

The Control group labelled phishing emails confidently 70% of the time (SD = 0.24), whereas the Experimental Group labelled them confidently 87% of the time (SD = 0.18). This difference is statistically significant (odds ratio = 0.34, p = 0.034). These findings confirm Hypothesis 2a. The use of the analysis module appears to have contributed to an increase in user confidence when labelling phishing emails.

6.2.2 Legitimate Emails

6.2.2.1 User Performance

According to Hypothesis 1b, it would be expected that the detection rate of the Experimental group for the legitimate emails would be higher when compared to the Control group. The control group were correct 86% of the time (SD = 0.11), compared to the

Experimental Group at 90% (SD = 0.09). This difference is not statistically significant (odds ratio = 0.70, p = 0.41). As such, there is there is currently insufficient evidence to confirm Hypothesis 1b.

70% and 45% of the Control and Experimental groups incorrectly marked email 4 as phishing, respectively. While this difference is not statistically significant (odds ratio = 0.357, p = 0.39), the level of incorrect labels is interesting. The email was external to the UoE, possessed a relatively long URL, and contained a financial incentive. This may explain the high rates of incorrect labelling.

The precision for legitimate emails of the Control and Experimental Group was 0.94 (SD = 0.08) and 0.96 (SD = 0.06), respectively. Both groups were often correct when they thought an email legitimate.

6.2.2.2 Confidence

According to Hypothesis 2b, it would be expected that the confidence of the Experimental group for the legitimate emails would be higher when compared to the Control group. The Control group labelled legitimate emails confidently 80% of the time (SD = 0.15), whereas the Experimental Group labelled them confidently 83% of the time (SD = 0.11). This difference is not statistically significant (odds ratio = 0.84, p = 0.62).

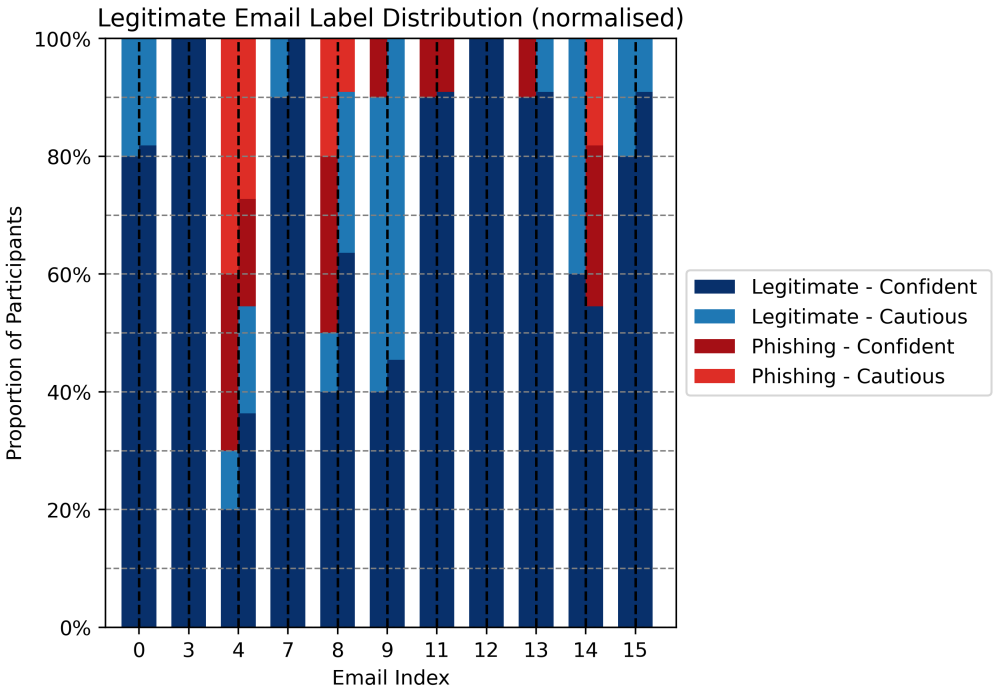


Figure 6.2: Labels assigned to each legitimate email in the email set by the participants. Each email has two bars: The left bar indicates the labels assigned by the control group, the right indicates the labels assigned by the experimental group.

6.2.3 Module Impact on Inbox Processing Time

The mean time the Control and Experimental group took to complete the task was 530 seconds (SD = 290) and 660 seconds (SD = 230), respectively. The Experimental group took $\approx 25\%$ longer, confirming Hypothesis 3. This may be as users took time to examine the information provided by the module when they analysed an email.

The mean difference in time that participants in the Control and Experimental group labelled each subsequent email on average was ≈ -2.6 seconds and ≈ -2.4 seconds, respectively, suggesting that users in both groups learned to label emails faster as they progressed through the list.

6.3 Summary

While the results showed that the use of the analysis module increased the confidence of user decisions when evaluating phishing emails, phishing detection rates did not increase by a significant amount. However, the experimental group did display increased precision in identifying phishing emails.

The analysis module was used on 31% of the emails viewed by the experimental group. When it was used, users complied with its classification 92% of the time, suggesting that users either trust the classification or invest less effort in reaching their own conclusions. The usage varied significantly from user to user, suggesting that some users relied upon the module for almost every email they saw, while others were more selective.

Participants in the Experimental group spent $\approx 25\%$ more time analysing the email set compared to the Control group. This is a significant increase in time - approximately 8 additional seconds per email. However, given that this was the participants' first exposure to a system of this kind, it may be that some of this added time is accounted for by the user learning to use the novel system.

The study did not reveal conclusive answers to many of the hypotheses stated. Further research is required to overcome this studies limitations and reach a conclusion upon whether the system improves user's phishing detection abilities. Regardless, some of the patterns of behaviours exhibited by users when employing the system have been identified which may inform said future research.

Chapter 7

Discussion and Evaluation

Given the combination of the damage that phishing attacks cause and the failures of current techniques in preventing users from interacting with emails unsafely, it is essential to investigate alternative solutions. Users have been shown to consistently focus on features of emails that are not reliable indicators of phishing when evaluating an email's legitimacy. Thus, this study aimed to evaluate the effectiveness of presenting users with additional information about reliable phishing indicators, as obtained and displayed by the automatic process employed by an email analysis module, implemented in the previous year. The study compared two different experimental conditions, wherein users were either shown generic, preventative advice or given access to the aforementioned analysis module. Through analysing the difference in user performance between these two groups, the study attempted to determine whether this information improved users' phishing detection rate and the magnitude of the improvement, among other statistics.

7.1 Interpretation of Results

Users with the ability to analyse are more precise and confident when dealing with phishing emails. The study found that presenting information such as that supplied from the email analysis module allows users to be more confident when labelling phishing emails, and they become more precise in detecting phishing emails. Despite the fact that the phishing detection rates between the two groups was not significantly different, the increased precision of the Experimental group suggests that the Control group was overly cautious, marking more emails as phishing than were present in the set. Thus, the hypothesis that users with the email analysis module are better at detecting phishing emails should not be totally discounted. These findings suggest there may be a positive relationship between module use and phishing detection. I speculate that the true detection rate of the two groups may be obscured by the Control groups imprecision - their detection rate may be inflated as they labelled more emails as phishing. However, this claim requires additional evidence to be substantiated.

The phishing detection rates of participants in both groups were approximately 90%. This value is notably higher than in similar studies [19, 91, 131], wherein most of the participants detected approximately half of the phishing emails. It may be that the

phishing emails in the set used for this study were too easily distinguishable, or that the participants were more adept at detecting phishing emails.

Users with the ability to analyse spend more time evaluating emails. Hypothesis 3 suggested that the Experimental group would spend more time analysing emails than the Control. Users with access to the analysis module were found to spend approximately 25% more time analysing the set of 16 emails when compared to those without, confirming this hypothesis. Considering the discussion in Section 2.2.2, this may pose a limitation to the module if used in a real-world setting. Security is a secondary goal for the user, and demanding they spend more time may dissuade users from using the analysis [43]. The added demand from the use of the module is unavoidable, and the increase in time users spend when using it is significant; this should be considered as a key statistic to measure in future research.

As the system is novel, this increased time may take into account the time it takes for a user to learn how to use the module. Both groups were found to label each subsequent label faster than those preceded it at a comparable rate. Given the small size of the email set, it is difficult to determine the effect of learning on module usage at this time.

Use of the analysis module varies widely. 4 of the 11 Experimental participants did not use the analysis module once, and each mislabelled at least one email. This is despite users being prompted to use the module in the instruction phase. This finding suggests that some participants may have felt confident in their ability to identify phishing emails without the assistance of the module, and/or they may not have fully understood its purpose or value. This hypothesis is supported by Kumaraguru et al., who claimed that users may be overconfident in their knowledge of phishing [55]. I found this result surprising, as I had expected users to analyse any email they could out of curiosity for the novel system. Instead, the mean usage rate was relatively low, at $\approx$31%.

The usage rate per participant varied widely. The mean number of times the module was used was 4.82, with an SD of 5.42, across the 16 emails. This large SD with respect to the mean suggests that some users relied upon the analysis module for the majority of the emails in the set, whereas others were more selective.

Users almost always comply with the analysis' classification. The exploratory analysis of the usage of the module discovered that, when used, the user agreed with its classification 92% of the time. This finding suggests that users place a high level of trust in the classification provided by the analysis module. It may be that users are relying on the module to be correct and are not investing the same amount of effort in reaching their own conclusions as the Control group. This hypothesis is supported by the discussion in Section 2.2.2. The specific analysis of email 14, the one email in the set incorrectly classified, showed this effect may be true even when the module is incorrect. This is despite the participants being warned in the instruction phase that the module may be incorrect. However, one email is insufficient in drawing a definitive conclusion on user behaviour when the module is incorrect.

The design of of the email analysis module's output may contribute to this effect. Tidwell discusses the guiding principles in designing single-page designs such as the one used to display the analysis [107]. Tidwell emphasises the importance of the *visual*

hierarchy. The visual hierarchy suggests users consider items at the top of designs as more important than items further down. The overall classification is placed at the top of the design, and is thus given a high level of importance respective to the rest of the information. This position of importance is in contrast to the aforementioned flaws in the overall classification of the module, discussed in Section 3.1. Despite the technical flaws behind how the classification is made, the design places it in an area of importance and users have been shown to rely upon that classification. As such, I suggest the design be altered to lower the importance of the classification, the classification process be improved upon to justify its position in the visual hierarchy, or be removed entirely.

7.2 Limitations

7.2.1 Participant Recruitment and Demographics

22 participants were recruited through an email sent to the undergraduate and post-graduate Master of Science student mailing lists within the University of Edinburgh's School of Informatics. This email can be seen in Appendix C. Table 7.1 summarises the demographics of this sample. This sample is not demographically representative, with the sample skewed towards male participants (64%) and the 18-24 age group (64%). Further, while it is not shown in Table 7.1, as the participants were recruited through the School of Informatics mailing list, they may possess an above average knowledge of cybersecurity concerns. This may be a reasonable explanation for the increased phishing detection rate of the participants. Together, these factors harm the generalisability of the results as they may not accurately reflect how the wider population would respond.

The most pertinent limitation of the study is that of the sample size (N=22) and number of phishing emails (5) used. Both are much less than that used in other studies [131, 14, 99, 122, 52]. As noted by Banerjee et al [6], this reduced the statistical power of the study, and may have led to Type II errors (false negatives) in the hypothesis the results of the study failed to confirm. Given the low sample size, few phishing emails, and increased phishing detection ability, there was a total of 11 incorrect labels in the 105 labels assigned to phishing emails, distributed between two groups. This amount of data is notably low, casting doubt on any conclusions reached using this data. A power analysis may be performed to determine the ideal sample size in future. Further, the lack of data prevented more advanced analyses, such as how the demographic sample varied in their usage and prediction rates; some of the demographic questions were answered by only 1 participant.

7.2.2 Website Environment and Methodology

This study suffers from the same limitation Zheng and Becker identified [131]. The website does not fully reflect an actual email inbox. Indeed, as I noted in more detail in Section 4.1, an embedded study, such as others conducted [14, 55, 99], wherein the email analysis module would be incorporated directly into user inboxes would result in a much improved study.

The data measured also could be improved upon. Evaluating a user's confidence when

	N	%		*N*	%
Current Level of Study			**Gender Identity**		
Pre-Honours (Years 1-2)	2	9%	Male	14	64%
Honours (Years 3-4)	4	18%	Female	8	36%
UG5	4	18%	Non-Binary	0	0%
PGT (Taught Post-Graduate)	5	23%	In Another Way	0	0%
PGR (Research Post-Graduate)	1	5%	Prefer Not to Say	0	0%
Other	6	27%			
Age			**Nationality**		
Under 18	0	0%	United Kingdom	12	55%
18-24	14	64%	Indonesia	3	14%
25-34	4	18%	United States of America	2	9%
35-44	3	14%	Poland	2	9%
45-54	0	0%	Turkey	1	5%
Over 54	1	5%	Greece	1	5%
			China	1	5%

Table 7.1: Demographics of the survey sample. Valid responses were received from a total of 22 respondents. All percentages shown to nearest integer.

they are labelling a phishing email is perhaps not as appropriate a measurement as evaluating confidence when the user *thought* the email was phishing - this would measure the user's confidence in their precision, which may be a more useful metric.

Participants in the study were able to access the analysis of each email within the time it took for their browser to receive the analysis from the website, as they had been analysed prior to the study as described in Section 5.1.4. This is not realistic, in a real-world setting the analysis will be conducted as and when a user asks for it. Users may have used the module more than they would in reality, as they don't have to wait for the process to occur. Introducing a delay in displaying the analysis to the user may mitigate this concern in future studies.

7.2.3 Email Analysis Module

The module itself possesses multiple limitations, some of which were identified in the previous year [95].

A defence-aware attacker may be able to circumvent many of the email analysis module's detection heuristics. If an attacker knew, for instance, the list of phishing keywords it uses in its analysis, an attacker may simply not use those words. This lowers the likelihood of a correct classification, and, given users' trust in that classification, would be detrimental to its effectiveness.

The current study and the module it evaluated did not consider the concept of 'information overload', as described by O'Reilly [83]. O'Reilly found that giving a user too much information is more detrimental to the user's decision-making process than if they had too little. Thus, the module may consider being more strategic in the information is provides to users in future.

Chapter 8

Conclusion and Future Work

It is indisputable that phishing email attacks will continue to cause damage for the foreseeable future. It has been consistently shown that users incorrectly assess evidence within emails that can conclusively distinguish them as phishing. Further, users fail to fully engage themselves in the analysis process, leading to an insufficient examination of said evidence. The study conducted within this paper attempted to address these failings by providing users with that evidence automatically, in such a way that they could read and understand that information. This information was provided using an automatic process developed across this paper and its predecessor [95], which analyses a range of features within emails and presents it to the user in a readable form - a technique that is novel to the phishing prevention field. A website previously used in phishing research [131] was altered, improved upon, and tested with the purpose of employing it within a study that aimed to compares users' phishing susceptibility with and without the system described. This study found that presenting users with this information increased their confidence and precision with regards to phishing emails. This finding contributes to the ongoing efforts to enhance email security; by improving the ability of individuals to accurately assess the information within phishing emails, these emails may be more readily identified, and the damage they cause may be mitigated. However, the study's limitations precluded any further assertions from being conclusively drawn. Most pertinently, it was not determined whether phishing detection rates increased by a meaningful amount. A future study of the same methodology will be conducted with increased resources to address these limitations and evaluate whether this novel system will be a beneficial addition to the phishing defence framework. Nonetheless, this study encourages a range of future works. Through the creation of similar user-centred tools that enhance users' email evaluation processes, the field may develop more effective strategies for combating the persistent threat that phishing attacks pose.

8.1 Future Work

As noted in Chapter 4, this study serves as a pilot study for a study that will be conducted using the same website I modified in Chapter 5. Thus, this section will aim its recommendations toward that study in addition to other studies that may seek to

address the limitations of this study.

This study's limitations prevented conclusions from being reached on the hypotheses of the study, most pertinent of which was whether this information improved users' phishing detection rate. This should be a key objective of future research on the module.

The usage of the email analysis module was shown to vary widely. Some participants used the module regularly, while others were more selective. It may be that some users rely on the module to analyse the majority of the emails they see, while others use it only when unsure. Future work should look to understand whether this, or another reason, is behind this wide range of usage. Further, some participants did not use the analysis at all. Further investigation is required to better understand the reasons behind this behaviour. Think-Aloud testing may be one approach to this investigation [49, 65], with the aim of gathering the user's thoughts to determine whether they never reference the module at all, or if they don't feel they need it. Another finding was that users spend much more time when they analysed emails, and that this time decreased as the users observed each subsequent email. Determining whether the amount of time spent analysing each email remains consistent or decreases as users become more familiar with the analysis may be an objective of future work. Furthermore, it would be beneficial to investigate whether the module's information can be presented more efficiently to lower this added time.

Including the demographics of the sample in the analyses may be beneficial. Zheng and Becker performed hierarchical linear regressions based off of their demographic data [131], which may be a possible extension to the methodology used in this paper. Further, including further demographic questions, such as allowing the users to self-report their knowledge of phishing, could provide further insights into what factors influence the usage of the module.

Other research may look to develop similar systems that supply users with additional information. The process used to evaluate many of of the phishing indicators analysed by the email analysis module can be improved upon; future work may look to refine the email analysis module by improving these indicators and/or incorporate more indicators or machine learning techniques. Additionally, different systems may be developed that perform a similar function. They could explore different types of information presentation; given previous research, this may include visual cues [16, 66, 87] or interactive elements [5], to facilitate user understanding and decision-making.

Bibliography

[1] Sara Albakry, Kami Vaniea, and Maria K. Wolters. *What is This URL's Destination? Empirical Evaluation of Users' URL Reading*, page 1–12. Association for Computing Machinery, New York, NY, USA, April 2020.

[2] Zainab Alkhalil, Chaminda Hewage, Liqaa Nawaf, and Imtiaz Khan. Phishing Attacks: A Recent Comprehensive Study and a New Anatomy. *Frontiers in Computer Science*, 3, 2021.

[3] Ammar Almomani, B. B. Gupta, Samer Atawneh, A. Meulenberg, and Eman Almomani. A survey of phishing email filtering techniques. *IEEE Communications Surveys & Tutorials*, 15(4):2070–2090, 2013.

[4] Kholoud Althobaiti, Adam DG Jenkins, and Kami Vaniea. A case study of phishing incident response in an educational organization. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–32, 2021.

[5] Kholoud Althobaiti, Kami Vaniea, and Serena Zheng. Faheem: Explaining URLs to people using a Slack bot. In *Symposium on Digital Behaviour Intervention for Cyber Security*, pages 1–8, April 2018. AISB 2018 Symposium on Swarm Intelligence & Evolutionary Computation, AISB 2018.

[6] Amitav Banerjee, UB Chitnis, SL Jadhav, JS Bhawalkar, and S Chaudhury. Hypothesis testing, type i and type ii errors. *Industrial psychiatry journal*, 18(2):127, 2009.

[7] Joe A. J. Beaumont. Spear Phishing - What It Is And How To Prevent It. URL: `https://www.bulletproof.co.uk/blog/what-is-spear-phishing`, May 2021. Online; Accessed 08-April-2023.

[8] André Bergholz, Gerhard Paaß, Frank Reichartz, Siehyun Strobel, and Schloß Birlinghoven. Improved Phishing Detection Using Model-based Features. In *Fifth Conference on Email and Anti-Spam, CEAS*, 2008.

[9] Tim Berners-Lee, Roy T. Fielding, and Larry M Masinter. Uniform Resource Identifier (URI): Generic Syntax. RFC 3986, January 2005.

[10] Al Bessey, Ken Block, Ben Chelf, Andy Chou, Bryan Fulton, Seth Hallem, Charles Henri-Gros, Asya Kamsky, Scott McPeak, and Dawson Engler. A few billion lines of code later: Using static analysis to find bugs in the real world. *Commun. ACM*, 53(2):66–75, feb 2010.

[11] BitBar. Browser & Mobile Testing for Apps. URL: https://smartbear.com/product/bitbar/. Online; Accessed 14-April-2023.

[12] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in neural information processing systems*, 2020.

[13] Can I Use. Support tables for HTML5, CSS3, etc. URL: https://caniuse.com/ciu/index. Online; Accessed: 16-April-2023.

[14] Deanna Caputo, Shari Pfleeger, Jesse Freeman, and M.Eric Johnson. Going Spear Phishing: Exploring Embedded Training and Awareness. *Security & Privacy, IEEE*, 12:28–38, January 2014.

[15] Kang Leng Chiew, Kelvin Sheng Chek Yong, and Choon Lin Tan. A survey of phishing attacks: Their types, vectors and technical approaches. *Expert Systems with Applications*, 106:1–20, 2018.

[16] Neil Chou, Robert Ledesma, Yuka Teraguchi, and John Mitchell. Client-Side Defense Against Web-Based Identity Theft. In *Network and Distributed System Security Symposium*, January 2004.

[17] Cofense. Annual Report 2021. URL: https://cofense.com/wp-content/uploads/2021/02/cofense-annual-report-2021.pdf. Online; Accessed 14-April-2023.

[18] Office 365 Developers. Outlook Add-ins Overview - Office Add-ins. URL: https://docs.microsoft.com/en-us/office/dev/add-ins/outlook/outlook-add-ins-overview. Online; Accessed 08-April-2023.

[19] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why Phishing Works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, page 581–590, New York, NY, USA, April 2006. Association for Computing Machinery.

[20] Daniel Eran Dilger. Apple apparently kills Windows PC support in Safari 6.0. URL: https://appleinsider.com/articles/12/07/25/apple_kills_windows_pc_support_in_safari_60, July 2012. Online; Accessed 14-April-2023.

[21] Agari Cyber Intelligence Division. 2021 Email Fraud and Identity Deception Trends. Technical report, Agari, 2021.

[22] DMARC.org. DMARC Overview. URL: https://dmarc.org/2022/01/dmarc-announced-ten-years-ago/. Online; Accessed 08-April-2023.

[23] DMARC.org. Statistics – DMARC. URL: https://dmarc.org/stats/dmarc. Online; Accessed 14-April-2023.

[24] DNSBL Information. Spam Database and Blacklist Check. URL: https://www.dnsbl.info/. Online; Accessed: 14-April-2023.

[25] Julie S Downs, Mandy Holbrook, and Lorrie Faith Cranor. Behavioral response to phishing risk. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, pages 37–44, 2007.

[26] Julie S. Downs, Mandy B. Holbrook, and Lorrie Faith Cranor. Decision Strategies and Susceptibility to Phishing. In *Proceedings of the Second Symposium on Usable Privacy and Security*, SOUPS '06, page 79–90, New York, NY, USA, 2006. Association for Computing Machinery.

[27] Ericka Chickowski. SOC Prevalence and Maturity on the Rise. URL: https://businessinsights.bitdefender.com/soc-prevalence-and-maturity-on-the-rise, November 2019. Online; Accessed 14-April-2023.

[28] Fastmail. Email Standards. URL: https://www.fastmail.help/hc/en-us/articles/1500000278382-Email-standards. Online; Accessed: 09-April-2023.

[29] UK Government Department for Digital, Culture, Media and Sport. Cyber Security Breaches Survey 2021. Technical report, March 2021.

[30] Anjuli Franz, Verena Zimmermann, Gregor Albrecht, Katrin Hartwig, Christian Reuter, Alexander Benlian, and Joachim Vogt. SoK: Still Plenty of Phish in the Sea — A Taxonomy of User-Oriented Phishing Interventions and Avenues for Future Research. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 339–358. USENIX Association, August 2021.

[31] Ned Freed and Nathaniel S. Borenstein. Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies. Request for Comments RFC 2045, Internet Engineering Task Force, November 1996.

[32] Jeremy Fuchs. Email Security Can Be Time Consuming. We Quantified The Exact Amount. URL: https://www.avanan.com/blog/email-security-can-be-time-consuming.-we-quantified-just-how-much, 2021. Online; Accessed 08-April-2023.

[33] Erich Gamma, Richard Helm, Ralph Johnson, Ralph E Johnson, and John Vlissides. *Design patterns: elements of reusable object-oriented software*. Pearson Deutschland GmbH, 1995.

[34] Greg Brockman, Atty Eleti, Elie Georges, Joanne Jang, Logan Kilpatrick, Rachel Lim, Luke Miller, Michelle Pokrass. Introducing ChatGPT and Whisper APIs. URL: https://openai.com/blog/introducing-chatgpt-and-whisper-apis. Online; Accessed 14-April-2023.

[35] Nielsen Norman Group. 10 Usability Heuristics for User Interface Design. URL: https://www.nngroup.com/articles/ten-usability-heuristics. Online; Accessed 13-April-2023.

[36] Nielsen Norman Group. Jakob Nielsen, Ph.D. and Principal at Nielsen Norman Group. URL: https://www.nngroup.com/people/jakob-nielsen. Online; Accessed 15-April-2023.

[37] Nielsen Norman Group. Usability 101: Introduction to Usability. URL: https://www.nngroup.com/articles/usability-101-introduction-to-usability. Online; Accessed 15-April-2023.

[38] The Radicati Group. Email Statistics, 2019-2023. Technical report, The Radicati Group, 2019.

[39] Frank E Grubbs. Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, pages 27–58, 1950.

[40] Shuang Hao, Nick Feamster, and Ramakant Pandrangi. Monitoring the Initial DNS Behavior of Malicious Domains. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, page 269–278, New York, NY, USA, November 2011. Association for Computing Machinery.

[41] Brynne Harrison, Elena Svetieva, and Arun Vishwanath. Individual processing of phishing emails. *Online Information Review*, 40:265–281, April 2016.

[42] Cormac Herley. So Long, And No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. In *Proceedings of the 2009 Workshop on New Security Paradigms Workshop*, NSPW '09, page 133–144, New York, NY, USA, January 2009. Association for Computing Machinery.

[43] Cormac Herley. More is not the answer. *IEEE Security and Privacy magazine*, January 2014.

[44] Adele Howe, Indrajit Ray, Mark Roberts, Malgorzata Urbanska, and Zinta Byrne. The psychology of security for the home computer user. *Proceedings - IEEE Symposium on Security and Privacy*, pages 209–223, 05 2012.

[45] Joel Hruska. Equifax Sent Customers to a Phishing Site, Hacked Months Earlier. *Extreme Tech*, September 2017.

[46] Internet Engineering Task Force. RFCs. URL: https://www.ietf.org/standards/rfcs/. Online; Accessed: 08-April-2023.

[47] Iulia Ion, Rob Reeder, and Sunny Consolvo. "...No one Can Hack My Mind": Comparing expert and Non-Expert security practices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 327–346, Ottawa, July 2015. USENIX Association.

[48] Markus Jakobsson, Alex Tsow, Ankur Shah, Eli Blevis, and Youn-Kyung Lim. What instills trust? a qualitative study of phishing. In *Financial Cryptography and Data Security: 11th International Conference, FC 2007, and 1st International*

Workshop on Usable Security, USEC 2007, Scarborough, Trinidad and Tobago, February 12-16, 2007. Revised Selected Papers 11, pages 356–361. Springer, 2007.

[49] Christopher J Johnstone, Nicole A Bottsford-Miller, and Sandra J Thompson. Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and english language learners. technical report 44. *National Center on Educational Outcomes, University of Minnesota*, 2006.

[50] Mahmoud Khonji, Youssef Iraqi, and Andrew Jones. Phishing Detection: A Literature Survey. *IEEE Communications Surveys and Tutorials*, 15:2091–2121, 2013.

[51] Iacovos Kirlappos and Angela Sasse. Security Education against Phishing: A Modest Proposal for a Major Rethink. *IEEE Security & Privacy*, 10:24–32, March 2012.

[52] Sabina Kleitman, Marvin KH Law, and Judy Kay. It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling. *PloS one*, 13(10):e0205089, 2018.

[53] Dr. John C. Klensin. Simple Mail Transfer Protocol. Technical Report RFC 5321, Internet Engineering Task Force, October 2008.

[54] Graham Klyne. Message Headers. URL: `https://www.iana.org/assignments/message-headers/message-headers.xhtml`, February 2022. Online; Accessed 08-April-2023.

[55] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Cranor, Jason Hong, and Elizabeth Nunge. Protecting People from Phishing: The Design and Evaluation of an Embedded Training Email System. *Conference on Human Factors in Computing Systems - Proceedings*, pages 905–914, April 2007.

[56] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)*, 10(2):1–31, 2010.

[57] Neil Kumaran. Spam Does Not Bring us Joy. URL: `https://cloud.google.com/blog/products/g-suite/ridding-gmail-of-100-million-more-spam-messages-with-tensorflow`, February 2021. Online; Accessed 08-April-2023.

[58] Pandove Kunal, Jindal Amandeep, and Kumar Rajinder. Email Spoofing. *International Journal of Computer Applications*, 5, August 2010.

[59] Elmer Lastdrager. Achieving a consensual definition of phishing based on a systematic review of the literature. *Crime Science*, 3(1):1–10, 2014.

[60] Elmer Lastdrager, Inés Carvajal Gallardo, Pieter Hartel, and Marianne Junger. How Effective is Anti-Phishing Training for Children? In *Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security*, SOUPS '17, page 229–239, USA, July 2017. USENIX Association.

[61] Page Laubheimer. Sticky Headers: 5 Ways to Make Them Better. URL:https://www.nngroup.com/articles/sticky-headers/. Online; Accessed 11-April-2023.

[62] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. A Research-oriented Top Sites Ranking Hardened Against Manipulation - Tranco. URL: https://tranco-list.eu/. Online; Accessed 08-April-2023.

[63] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings 2019 Network and Distributed System Security Symposium*, San Diego, CA, 2019. Internet Society.

[64] Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*, volume 3. Springer, 2005.

[65] Clayton Lewis, John Rieman, and Task-Centered User Interface Design. Task-Centered User Interface Design: A Practical Introduction. *University of Colorado, Boulder, Department of Computer Science*, page 20, 1993.

[66] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. Does Domain Highlighting Help People Identify Phishing Sites? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2075–2084, Vancouver BC Canada, May 2011. ACM.

[67] Eric Lipton, David E. Sanger, and Scott Shane. The Perfect Weapon: How Russian Cyberpower Invaded the U.S. *The New York Times*, December 2016.

[68] Andrew J Lohn and Krystal Alex Jackson. Will ai make cyber swords or shields: A few mathematical models of technological progress, 2022.

[69] Xin Robert Luo, Wei Zhang, Stephen Burd, and Alessandro Seazzu. Investigating phishing victimization with the heuristic–systematic model: A theoretical framework and an exploration. *Computers & Security*, 38:28–38, 2013.

[70] Bella Martin, Bruce Hanington, and Bruce M Hanington. *Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions*. Rockport Pub, 2012.

[71] John H McDonald. *Handbook of Biological Statistics*, volume 3. Sparky House Publishing, Baltimore, Maryland, 2014.

[72] Alexey Melnikov and Barry Leiba. Internet Message Access Protocol (IMAP) - Version 4rev2. Technical Report 9051, Internet Engineering Task Force, August 2021.

[73] Bertrand Meyer. *Object-oriented software construction*, volume 2. Prentice hall Englewood Cliffs, 1997.

[74] Microsoft 365. Microsoft Outlook. URL: https://www.microsoft.com/en-gb/microsoft-365/outlook/email-and-calendar-software-microsoft-outlook. Online; Accessed: 08-April-2023.

[75] Mozilla Developer Network Web Documentation. <iframe>: The Inline Frame element - HTML: HyperText Markup Language. URL: https://developer.mozilla.org/en-US/docs/Web/HTML/Element/iframe/. Online; Accessed: 08-April-2023.

[76] Mozilla Developer Network Web Documentation. What is JavaScript? - Learn Web Development. URL: https://developer.mozilla.org/en-US/docs/Learn/JavaScript/First_steps/What_is_JavaScript, March 2022. Online; Accessed: 17-April-2023.

[77] Mozilla Developer Network Web Documentation. POST - HTTP. URL:https://developer.mozilla.org/en-US/docs/Web/HTTP/Methods/POST/, April 2023. Online; Accessed 11-April-2023.

[78] Paula M.W. Musuva, Katherine W. Getao, and Christopher K. Chepken. A new approach to modelling the effects of cognitive processing and threat detection on phishing susceptibility. *Computers in Human Behavior*, 94:154–175, 2019.

[79] Jakob Nielsen. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, page 152–158, New York, NY, USA, 1994. Association for Computing Machinery.

[80] Federal Bureau of Investigation. 2020 Internet Crime Report. Technical report, Federal Bureau of Investigation, 2020.

[81] Office 365 Team. Microsoft Forms—a new formative assessment and survey tool in Office 365 Education - Office Blogs. URL:https://web.archive.org/web/20170623024338/https://blogs.office.com/2016/06/20/microsoft-forms-a-new-formative-assessment-and-survey-tool-in-office-365-education/, June 2016. Online; Accessed 10-April-2023. Archived 23-June-2017.

[82] Optimizely. User Flow. URL: https://www.optimizely.com/optimization-glossary/user-flow/. Online; Accessed: 09-April-2023.

[83] Charles A. O'Reilly. Individuals and Information Overload in Organizations: Is More Necessarily Better? *The Academy of Management Journal*, 23(4):684–696, 1980.

[84] Overleaf. Online LaTeX Editor. URL: https://www.overleaf.com/about/. Online; Accessed: 16-April-2023.

[85] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, and Cate Jerram. Phishing for the truth: A scenario-based experiment of users' behavioural response to emails. In Lech J. Janczewski, Henry B. Wolfe, and Sujeet Shenoi, editors, *Security and Privacy Protection in Information Processing Systems*, pages 366–378, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[86] William D. Perreault. Controlling order-effect bias. *The Public Opinion Quarterly*, 39(4):544–551, 1975.

[87] Justin Petelka, Yixin Zou, and Florian Schaub. Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings. *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15, April 2019.

[88] Christopher J. Prom. *Preserving Email*. Number 11-01 in DPC Technology Watch Report. Digital Preservation Coalition, December 2011.

[89] ProofPoint. State Of The Phish – An In-depth Look at User Awareness, Vulnerability and Resilience. Technical report, ProofPoint Inc., 2022.

[90] Pete Resnick. Internet Message Format. Request for Comments RFC 5322, Internet Engineering Task Force, October 2008. Num Pages: 57.

[91] Stefan A. Robila and James W. Ragucci. Don't Be a Phish: Steps in User Education. In *Proceedings of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education*, ITICSE '06, page 237–241, New York, NY, USA, January 2006. Association for Computing Machinery.

[92] Roman Dedenok. Spam mail with scammers' phone numbers. URL: `https://www.kaspersky.co.uk/blog/spam-with-vishing-phone-numbers/23163`, August 2021. Online; Accessed 15-April-2023.

[93] Maddie Rosenthal. Phishing Statistics (Updated 2022) - 50+ Important Phishing Stats. URL: `https://www.tessian.com/blog/phishing-statistics-2020/`, January 2022. Online; Accessed 08-April-2023.

[94] Tara Seals. 84% of Phishing Sites Last for Less Than 24 Hours. URL: `https://www.infosecurity-magazine.com/news/84-of-phishing-sites-last-for-less/`, December 2016. Online; Accessed 08-April-2023.

[95] Sean Strain. Automatically Generating Contextualised Responses to Phishing Reports. *University of Edinburgh, School of Informatics*, 2022.

[96] Selenium. Web Browser Automation. URL: `https://www.selenium.dev`. Online; Accessed 14-April-2023.

[97] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[98] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Cranor, Jason Hong, and Elizabeth Nunge. Anti-Phishing Phil: The design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, volume 229, pages 88–99, January 2007.

[99] Hossein Siadati, Sean Palka, Avi Siegel, and Damon McCoy. Measuring the effectiveness of embedded phishing exercises. In *10th USENIX Workshop on Cyber Security Experimentation and Test (CSET 17)*, Vancouver, BC, August 2017. USENIX Association.

[100] Stu Sjouwerman. Phone Number Only Phishing on the Rise. URL: https://blog.knowbe4.com/phone-number-only-phishing-on-the-rise, February 2022. Online; Accessed 15-April-2023.

[101] Rebecca Smith. How a U.S. Utility Got Hacked. *Wall Street Journal*, December 2016.

[102] Eden Spivak. The Iterative Design Process: A Full Guide for UX Designers, February 2021.

[103] Statcounter. Browser Market Share Worldwide. URL: https://gs.statcounter.com/browser-market-share. Online; Accessed 14-April-2023.

[104] Stephanie Carruthers, Camille Singleton and Charles DeBeck. Why Phishing Is Still the Top Attack Method. URL: https://securityintelligence.com/posts/why-phishing-still-top-attack-method-2, July 2022. Online; Accessed 15-April-2023.

[105] Michelle Steves, Kristen Greene, and Mary Theofanos. Categorizing human phishing difficulty: a Phish Scale. *Journal of Cybersecurity*, 6(1), September 2020. _eprint: https://academic.oup.com/cybersecurity/article-pdf/6/1/tyaa009/33746006/tyaa009.pdf.

[106] Sustainability of Digital Formats. Email (Electronic Mail Format). URL: https://www.loc.gov/preservation/digital/formats/fdd/fdd000388.shtml, April 2014. Online; Accessed: 08-April-2023.

[107] Jenifer Tidwell. *Designing Interfaces: Patterns for Effective Interaction Design*. O'Reilly Media Inc., November 2005.

[108] Miles Tracy, Wayne Jansen, Karen Scarfone, and Jason Butterfield. Guidelines on Electronic Mail Security. URL: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-45ver2.pdf, 2007. Online; Accessed: 18-April-2023.

[109] TrendMicro. Spear-Phishing Email: Most Favored APT Attack Bait. URL: https://ddos.inforisktoday.com/whitepapers/spear-phishing-email-most-favored-apt-attack-bait-w-664, December 2012. Online; Accessed 15-April-2023.

[110] John W Tukey et al. *Exploratory data analysis*, volume 2. Reading, MA, 1977.

[111] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458, 1981.

[112] Steffen Ullrich. Breaking DKIM - on Purpose and by Chance. URL: https://noxxi.de/research/breaking-dkim-on-purpose-and-by-chance.html/, October 2017. Online; Accessed 18-April-2023.

[113] René van Bavel, Nuria Rodríguez-Priego, José Vila, and Pam Briggs. Using protection motivation theory in the design of nudges to improve online security behavior. *International Journal of Human-Computer Studies*, 123:29–39, 2019.

[114] Guido van Rossum, Barry Warsaw, and Nick Coghlan. PEP 8 – Style Guide for Python Code. Technical report, Python, 2001.

[115] Verizon. 2019 Data Breach Investigations Report. Technical report, Verizon, 2019.

[116] Verizon. 2021 Data Breach Investigations Report. Technical report, Verizon, 2021.

[117] Verizon. 2022 Data Breach Investigations Report. Technical report, Verizon, 2022.

[118] Arun Vishwanath, Tejaswini Herath, Rui Chen, Jingguo Wang, and H Raghav Rao. Why do people get phished? testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, 51(3):576–586, 2011.

[119] Melanie Volkamer, Karen Renaud, and Paul Gerber. Spot the Phish by Checking the Pruned URL. *Information & Computer Security*, 24(4):372–385, January 2016. Publisher: Emerald Group Publishing Limited.

[120] Vue. Vue.js - The Progressive JavaScript Framework. URL: `https://vuejs.org/`. Online; Accessed: 08-April-2023.

[121] Jingguo Wang, Tejaswini Herath, Rui Chen, Arun Vishwanath, and H. Raghav Rao. Research article phishing susceptibility: An investigation into the processing of a targeted spear phishing email. *IEEE Transactions on Professional Communication*, 55(4):345–362, 2012.

[122] Rick Wash. How Experts Detect Phishing Scam Emails. *Proceedings of the ACM on Human-Computer Interaction*, 4, October 2020.

[123] Rick Wash, Norbert Nthala, and Emilee Rader. Knowledge and Capabilities that Non-Expert Users Bring to Phishing Detection. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 377–396. USENIX Association, August 2021.

[124] Web Hypertext Application Technology Working Group. The HTML Standard. URL: `https://html.spec.whatwg.org/multipage/introduction.html/`. Online; Accessed 18-April-2023.

[125] Alma Whitten and J. D. Tygar. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *Proceedings of the 8th Conference on USENIX Security Symposium - Volume 8*, SSYM'99, page 14, USA, August 1999. USENIX Association.

[126] Who Called Me? Free Reverse Phone Lookup. URL: `https://whocalled.co.uk//`, April 2023. Online; Accessed: 16-April-2023.

[127] Wikipedia contributors. Comparison of Email Clients. URL: `https://en.wikipedia.org/w/index.php?title=Comparison_of_email_clients&oldid=1077003658`, April 2023. Page Version ID: 1145564305. Online; Accessed 08-April-2023.

[128] World Wide Web Consortium (W3C). Web Content Accessibility Guidelines (WCAG) 2.1. URL: https://www.w3.org/TR/WCAG21/. Online; Accessed: 16-April-2023.

[129] Kinza Yasar. What is Software Testing? Definition, Types and Importance. URL:https://www.techtarget.com/whatis/definition/software-testing/. Online; Accessed 11-April-2023.

[130] Zeyu Zhang. Designing an Autoresponder for Phishing Email Reports. [Unpublished manuscript], 2021.

[131] Sarah Zheng and Ingolf Becker. Presenting Suspicious Details in User-Facing E-mail Headers Does Not Improve Phishing Detection. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 253–271, 2022.

Appendix A

Website Tests

A.1 Think-Aloud Test and Results

The following test follows the recommendations of Martin et al [70] and Lewis et al [65].

The two participants were asked to complete the following task: “Complete the labelling of all emails within the website”.

I sat next to the participant while they completed the task. I asked them to speak aloud their thoughts as they conducted the task. The participants were not recorded with audio or video equipment. I noted the relevant thoughts they verbalised as accurately to what was spoken as possible.

A.1.1 Participant One

I can classify an image [sic] *as legitimate many times even after I’ve already classified it as legitimate. Makes it hard to tell if I’ve already classified them.*

Refreshing the page completely breaks it (though that’s probably expected?).

The buttons for ‘get more information’ and then ’analyse’ is redundant. Why do I have to click twice when I could click just once for the same effect?

The button to resize the email is really small and the grey makes it hard to see.

The download file won’t open in Excel, is it supposed to?

All in all I think it looks really professional. It just needs a few changes here and there.

A.1.2 Participant Two

In the middle panel, the emails don’t always expand fully which is annoying, and you can’t see the scrollbar unless you go down to the bottom of the whole page.

I have to scroll back up every time I want to label an email and then back down again and it's actually starting to annoy me. How about having the header maybe be persistent and not scroll away when I scroll down? That might be a bit difficult to code though.

It doesn't give any visual markers to show me what I've marked good or bad, I don't know if that's part of the design but.

This email is literally like 10 or 20 pixels too small on the screen and the scrollbar is showing.

At this point I showed the participant they could resize the email using the resize button.

Okay that's way too small. I had no idea you could even do that.

I made it big! - The participant had used the resize tool to make the email so wide on the screen it wrapped around and was now displayed underneath the email list.

Yeah you need to change that resize stuff, like it needs to be more obvious and you need to limit the sizes.

A.2 Screenshot Tests

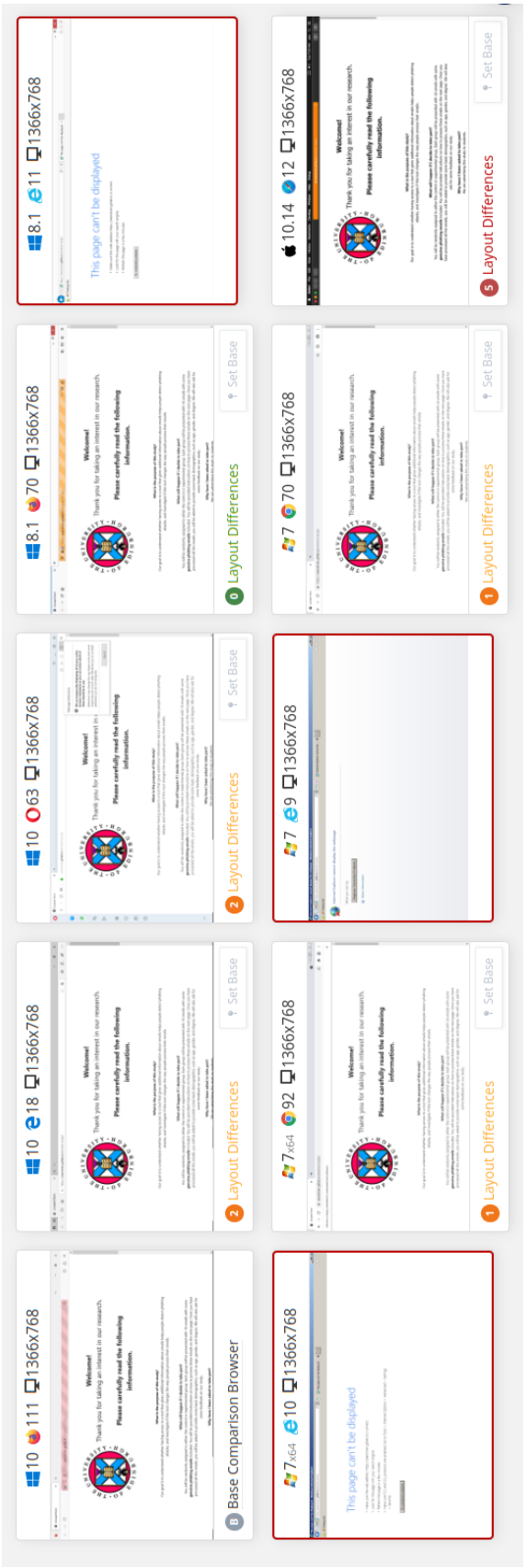
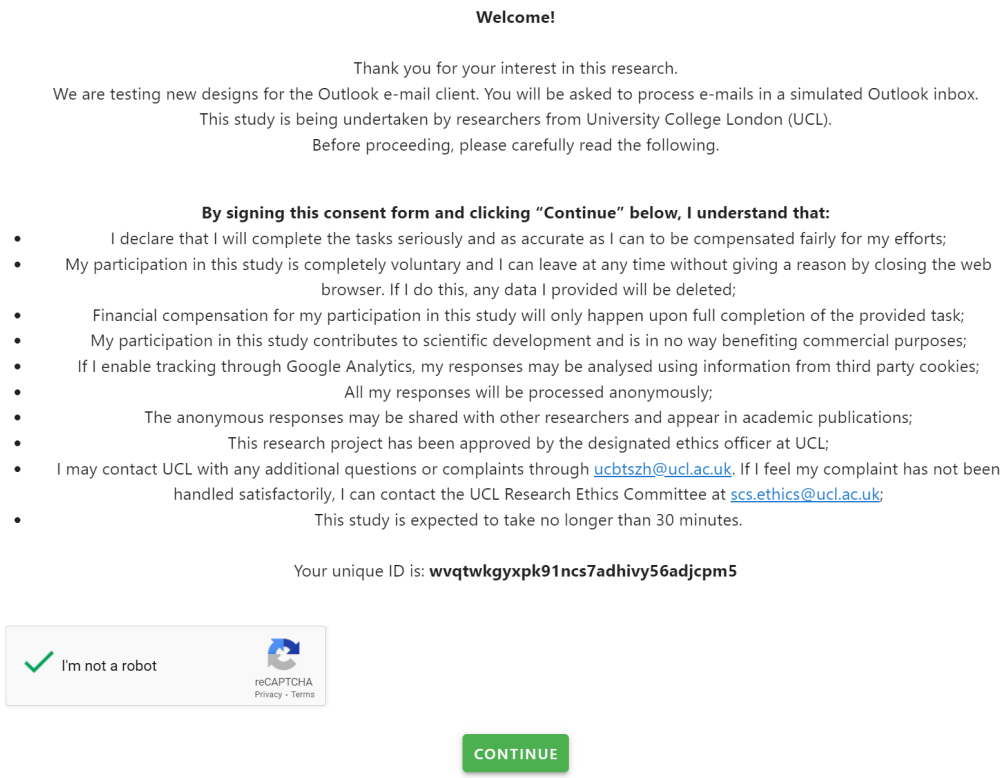


Figure A.1: Screenshot Test.

Appendix B

Study Website Design

B.1 Unmodified Website



Welcome!

Thank you for your interest in this research.
We are testing new designs for the Outlook e-mail client. You will be asked to process e-mails in a simulated Outlook inbox.
This study is being undertaken by researchers from University College London (UCL).
Before proceeding, please carefully read the following.

By signing this consent form and clicking "Continue" below, I understand that:

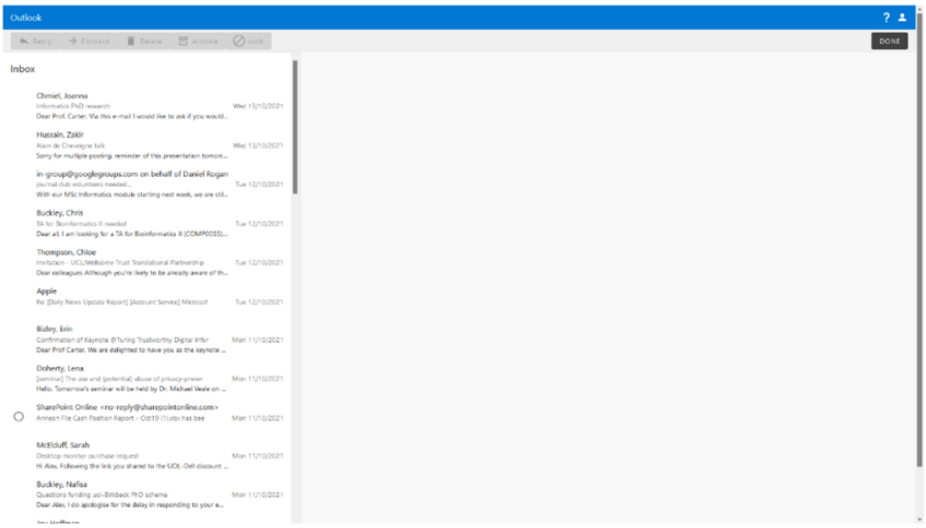
- I declare that I will complete the tasks seriously and as accurate as I can to be compensated fairly for my efforts;
- My participation in this study is completely voluntary and I can leave at any time without giving a reason by closing the web browser. If I do this, any data I provided will be deleted;
- Financial compensation for my participation in this study will only happen upon full completion of the provided task;
- My participation in this study contributes to scientific development and is in no way benefiting commercial purposes;
- If I enable tracking through Google Analytics, my responses may be analysed using information from third party cookies;
- All my responses will be processed anonymously;
- The anonymous responses may be shared with other researchers and appear in academic publications;
- This research project has been approved by the designated ethics officer at UCL;
- I may contact UCL with any additional questions or complaints through ucbtszh@ucl.ac.uk. If I feel my complaint has not been handled satisfactorily, I can contact the UCL Research Ethics Committee at scs.ethics@ucl.ac.uk;
- This study is expected to take no longer than 30 minutes.

Your unique ID is: **wvqtwkgyxpk91ncs7adhivy56adjcpm5**

Figure B.1: Unmodified Consent Page.

Imagine that you are a renowned professor named Alex Carter in the field of health informatics, working at University of London (UOL).

In this study, you will see a mock inbox with 47 e-mails adapted from actual e-mails received by academics.
Your task is to process these e-mails as if you received them in your role of Prof. Carter.



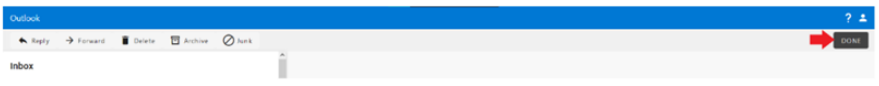
The inbox has similar functionalities as what you typically would find in an actual e-mail inbox.
To process an e-mail, click on an e-mail in the list on the left.
The full e-mail will be displayed on the right.

All available actions that you can take with the selected e-mail are displayed in the bar above the e-mails.

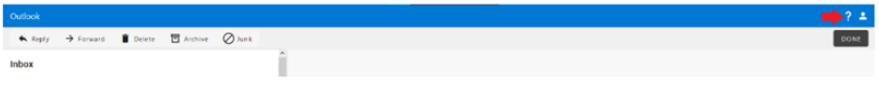
- **Reply**: allows you to reply to the sender of the e-mail.
- **Forward**: forward the e-mail to your executive assistant (EA), so the EA can take care of it. You do not need to type an extra message for this action.
- **Archive**: move the e-mail to an archive.
- **Delete**: delete the e-mail.
- **Junk**: report the e-mail as junk and move it to the junk folder.

Your task is to choose what you think is the **best applicable action** for each e-mail.
Please note that all links in e-mails have been disabled and that this task should not take more than 30-45 minutes.

When you are done processing all e-mails, click "DONE":



You can review these instructions at any time during the task through the "help" icon:



Click "START" below to go to the inbox.

START

Figure B.2: Unmodified Instruct Page.

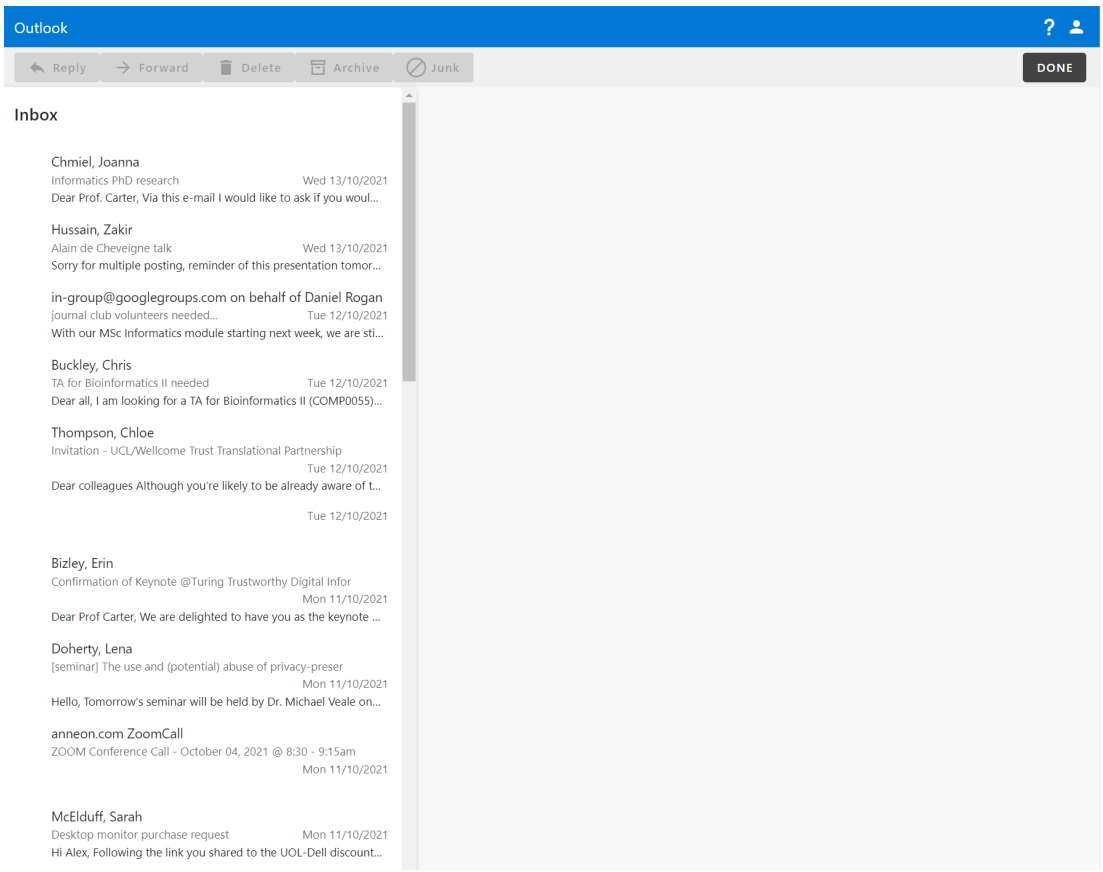
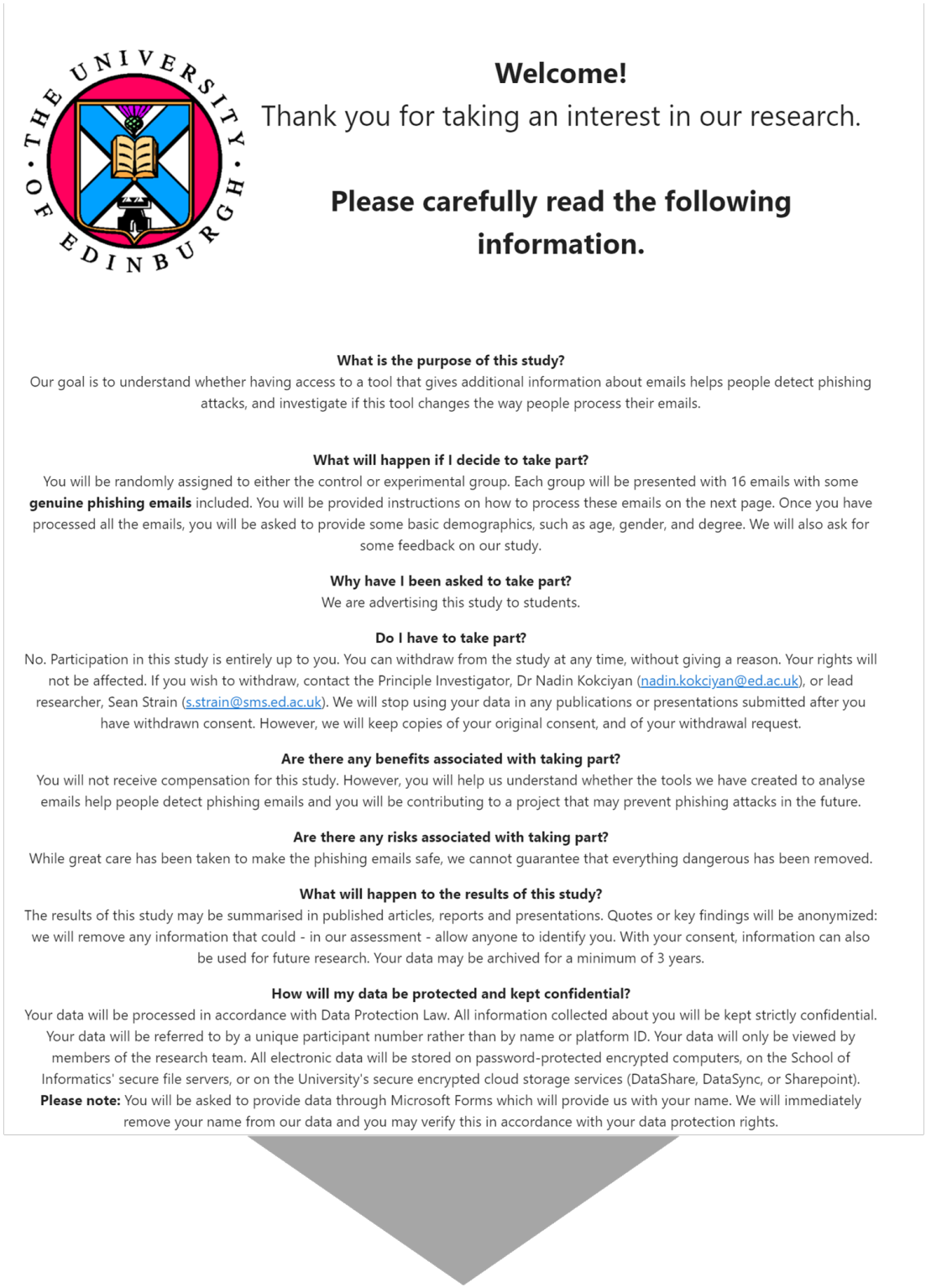


Figure B.3: Unmodified Task Page.

B.2 Modified Website as Described in Section 5.1



Welcome!

Thank you for taking an interest in our research.

Please carefully read the following information.

What is the purpose of this study?

Our goal is to understand whether having access to a tool that gives additional information about emails helps people detect phishing attacks, and investigate if this tool changes the way people process their emails.

What will happen if I decide to take part?

You will be randomly assigned to either the control or experimental group. Each group will be presented with 16 emails with some **genuine phishing emails** included. You will be provided instructions on how to process these emails on the next page. Once you have processed all the emails, you will be asked to provide some basic demographics, such as age, gender, and degree. We will also ask for some feedback on our study.

Why have I been asked to take part?

We are advertising this study to students.

Do I have to take part?

No. Participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. Your rights will not be affected. If you wish to withdraw, contact the Principle Investigator, Dr Nadin Kokciyan (nadin.kokciyan@ed.ac.uk), or lead researcher, Sean Strain (s.strain@sms.ed.ac.uk). We will stop using your data in any publications or presentations submitted after you have withdrawn consent. However, we will keep copies of your original consent, and of your withdrawal request.

Are there any benefits associated with taking part?

You will not receive compensation for this study. However, you will help us understand whether the tools we have created to analyse emails help people detect phishing emails and you will be contributing to a project that may prevent phishing attacks in the future.

Are there any risks associated with taking part?

While great care has been taken to make the phishing emails safe, we cannot guarantee that everything dangerous has been removed.

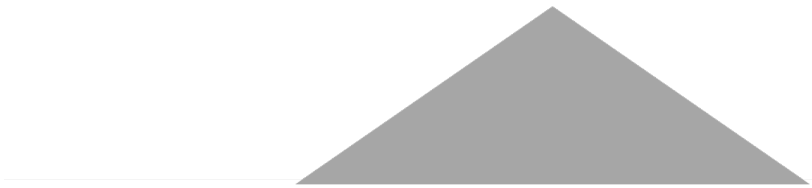
What will happen to the results of this study?

The results of this study may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: we will remove any information that could - in our assessment - allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a minimum of 3 years.

How will my data be protected and kept confidential?

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name or platform ID. Your data will only be viewed by members of the research team. All electronic data will be stored on password-protected encrypted computers, on the School of Informatics' secure file servers, or on the University's secure encrypted cloud storage services (DataShare, DataSync, or Sharepoint). **Please note:** You will be asked to provide data through Microsoft Forms which will provide us with your name. We will immediately remove your name from our data and you may verify this in accordance with your data protection rights.

Figure B.4: Modified Consent Page Part 1 of 2.



What are my data protection rights?

The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk. For general information about how we use your data, go to: edin.ac/privacy-research

Who is conducting this research and who can I contact about it?

This survey is conducted by researchers from School of Informatics, University of Edinburgh. If you have any further questions about the study, please contact the lead researcher, Sean Strain (s.strain@sms.ed.ac.uk), or his supervisor Dr Nadin Kokciyan (nadin.kokciyan@ed.ac.uk).

If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint. This study was certified according to the Informatics Ethics boards, with reference number 2022/82433.

Updated information

If the research project changes in any way, an updated Participant Information Sheet will be made available on: web.inf.ed.ac.uk/infweb/research/study-updates

Alternative formats

To request this document in an alternative format, such as large print or on coloured paper, please contact Sean Strain (s.strain@sms.ed.ac.uk).

Your unique ID is: **wwpdggkxsz4zni5ioazvqknbg3havkuo**

By clicking consent, you agree to all of the terms listed above.

I CONSENT

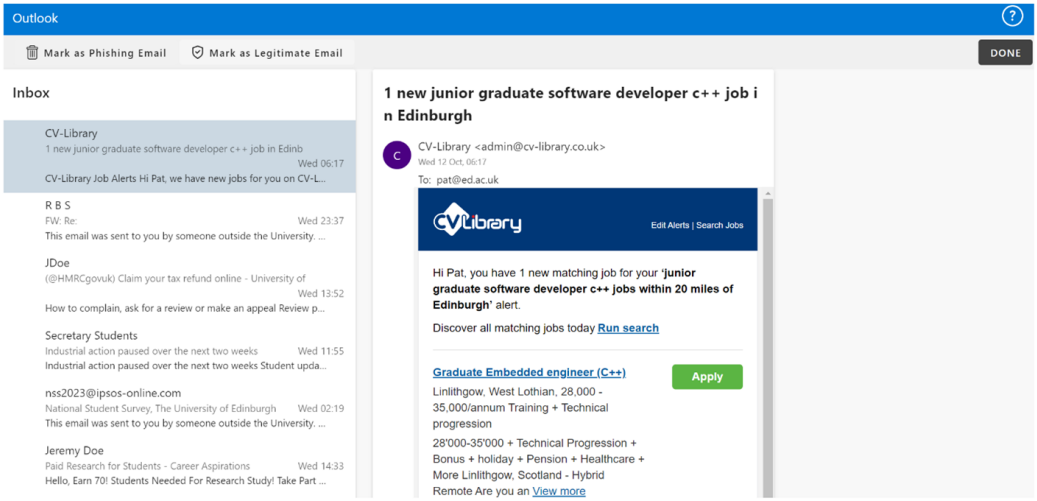
Figure B.5: Modified Consent Page Part 2 of 2.

Imagine you are a 5th year undergraduate masters student at the University of Edinburgh - named **Pat Smith**.

As a student you often receive emails from your university, but also from your personal accounts that you have registered with. This means your emails can range from generic university announcements to security alerts for your personal accounts.

Now imagine that you open your inbox and find 16 new emails from the day.
Your task is to process them all in your role as Pat Smith.

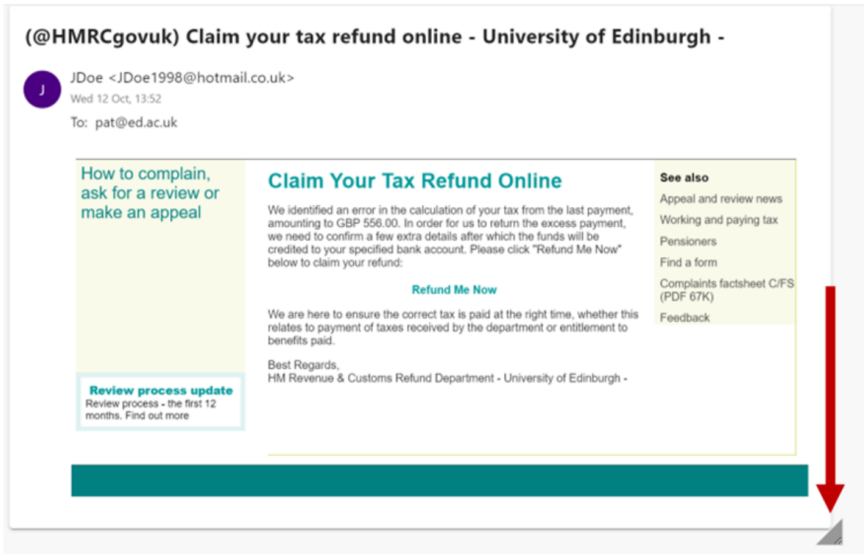
You will see a mock inbox like this:



You are able to click on the emails on the left to move between each email just as you would in a real inbox.

If the email is too small on your screen, you can use the resize tool at the bottom right to resize it.
That tool looks like this:

Figure B.6: Modified Instruct Page Part 1 of 4.



You are asked to complete one of the following steps for each email:

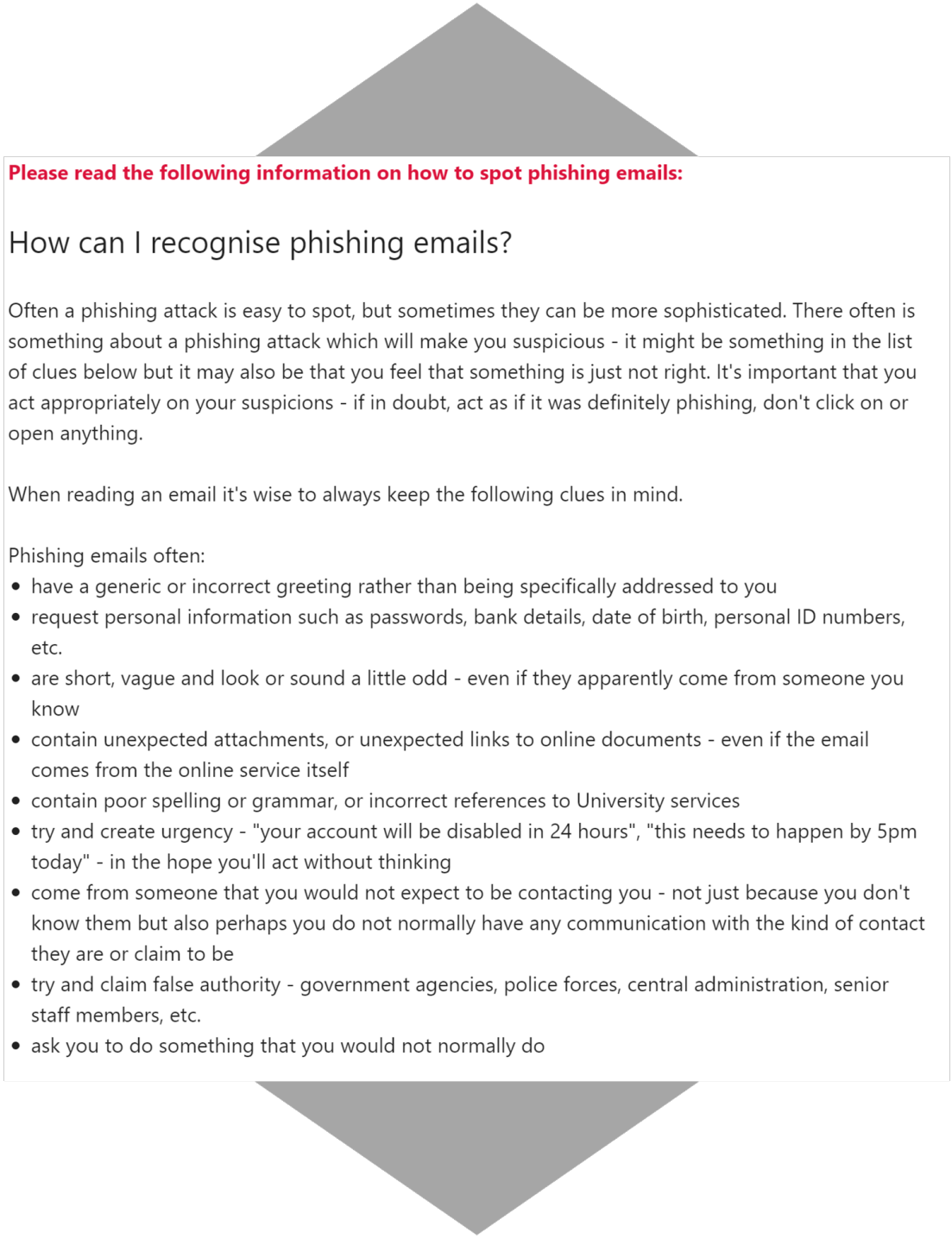
• **Mark as Phishing Email**: this option marks the email as phishing and will make the listing in the inbox red.
• **Mark as Legitimate Email**: this option marks the email as legitimate and will make the listing in the inbox green.

There are options within that allow you to say if you are unsure - just give your best guess.

You may go back and change your mind at any time. Simply click on the email again and choose a different action.

Some of the emails in your inbox are phishing emails. **Your job is to spot these emails and mark them accordingly.**

Figure B.7: Modified Instruct Page Part 2 of 4.



Please read the following information on how to spot phishing emails:

How can I recognise phishing emails?

Often a phishing attack is easy to spot, but sometimes they can be more sophisticated. There often is something about a phishing attack which will make you suspicious - it might be something in the list of clues below but it may also be that you feel that something is just not right. It's important that you act appropriately on your suspicions - if in doubt, act as if it was definitely phishing, don't click on or open anything.

When reading an email it's wise to always keep the following clues in mind.

Phishing emails often:

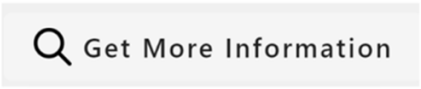
- have a generic or incorrect greeting rather than being specifically addressed to you
- request personal information such as passwords, bank details, date of birth, personal ID numbers, etc.
- are short, vague and look or sound a little odd - even if they apparently come from someone you know
- contain unexpected attachments, or unexpected links to online documents - even if the email comes from the online service itself
- contain poor spelling or grammar, or incorrect references to University services
- try and create urgency - "your account will be disabled in 24 hours", "this needs to happen by 5pm today" - in the hope you'll act without thinking
- come from someone that you would not expect to be contacting you - not just because you don't know them but also perhaps you do not normally have any communication with the kind of contact they are or claim to be
- try and claim false authority - government agencies, police forces, central administration, senior staff members, etc.
- ask you to do something that you would not normally do

Figure B.8: Modified Instruct Page Part 3A of 4- Control Group User Flow.

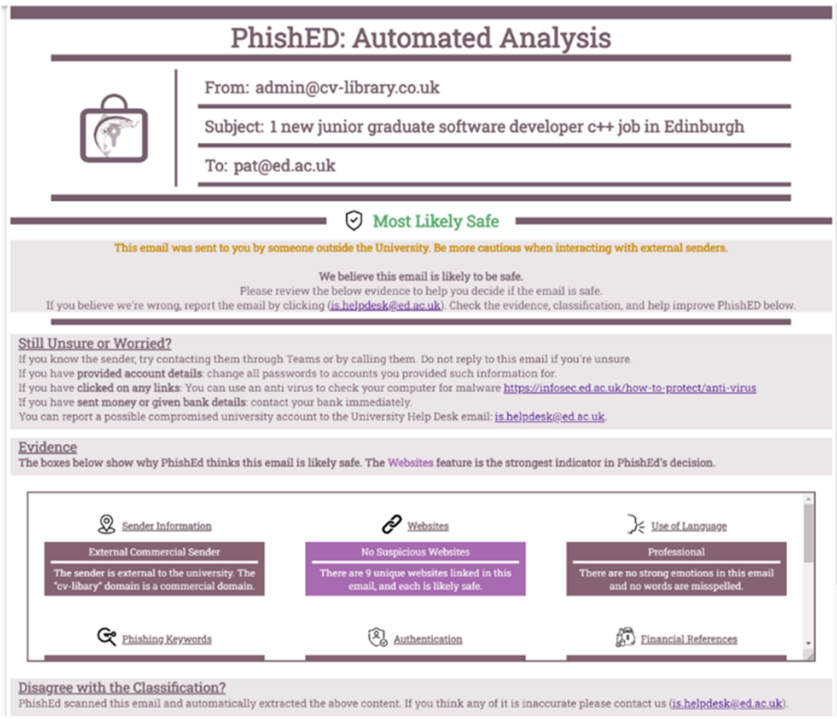
You also have access to a tool that analyses the emails for you.

This tool can be accessed through the **Get More Information** button. This will allow you to see additional information about the email which can help you decide if it is a phishing email or not. You are welcome to use this tool as much as you like.

The button looks like this:



You will then see a document appear that looks like this:



What information does the tool provide? How does the tool work?:

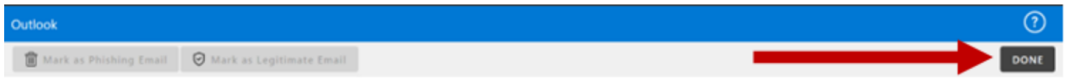
Each of the following emails you see in your inbox has been analysed by an automated script that evaluates the email for certain features that are common in phishing emails. For instance, it looks for malicious websites in the email or the use of words commonly found in phishing emails. A machine learning algorithm then uses these features to make a prediction about whether an email is a phishing email or not. The tool provides you with the results of this prediction and what features factored into its decision.

Please Note: this tool is sometimes wrong about it's overall estimation. It is meant to help you, not decide for you. There are both legitimate emails that are evaluated as phishing emails, and phishing emails that are evaluated as legitimate emails in this inbox. You should always use your own judgement when deciding if an email is legitimate or not.

Figure B.9: Modified Instruct Page Part 3B of 4- Experimental Group User Flow.

To process an email:

1. Click on an email in the list on the left. The full email will be displayed on the right.
2. **Choose one of the available actions** for the selected email from the bar displayed above the emails.
3. A secondary bar with buttons will appear. Click on the button that **best represents** your decision.
4. The email will be coloured red or green depending on your decision.
5. You are able to go back and change your mind at any point. Simply click on the email again and choose a different action.
6. A confirmation of your processing action will be shown in the bottom of the screen.
7. **When you have processed all emails, click "DONE":** - This will take you to our survey and download your results.

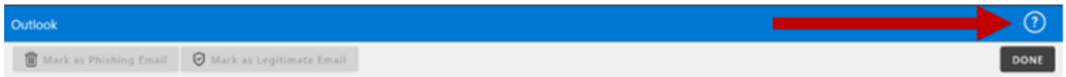


The results file will be of the ".xlsx" type, but your computer won't be able to read it. This is expected - we will change the file format to ".txt" upon downloading your results. It has to be '.xlsx' so that Microsoft Forms can accept it. You may look at the data by altering the file extension to ".txt" should you so wish.

Finally, a very short survey will be presented to you. Please complete this survey to submit your answers. This survey requires you to upload the file from your computer. You can find the file in the "downloads" folder.

All links in these emails have been altered to make them safe. This task should not take more than 5-10 minutes.

You can review these instructions at any time during the task through the "help" icon:



Click on the button below to confirm that you understood the task and are ready to start.

One last thing: Thank you!

I UNDERSTAND

Figure B.10: Modified Instruct Page Part 4 of 4.

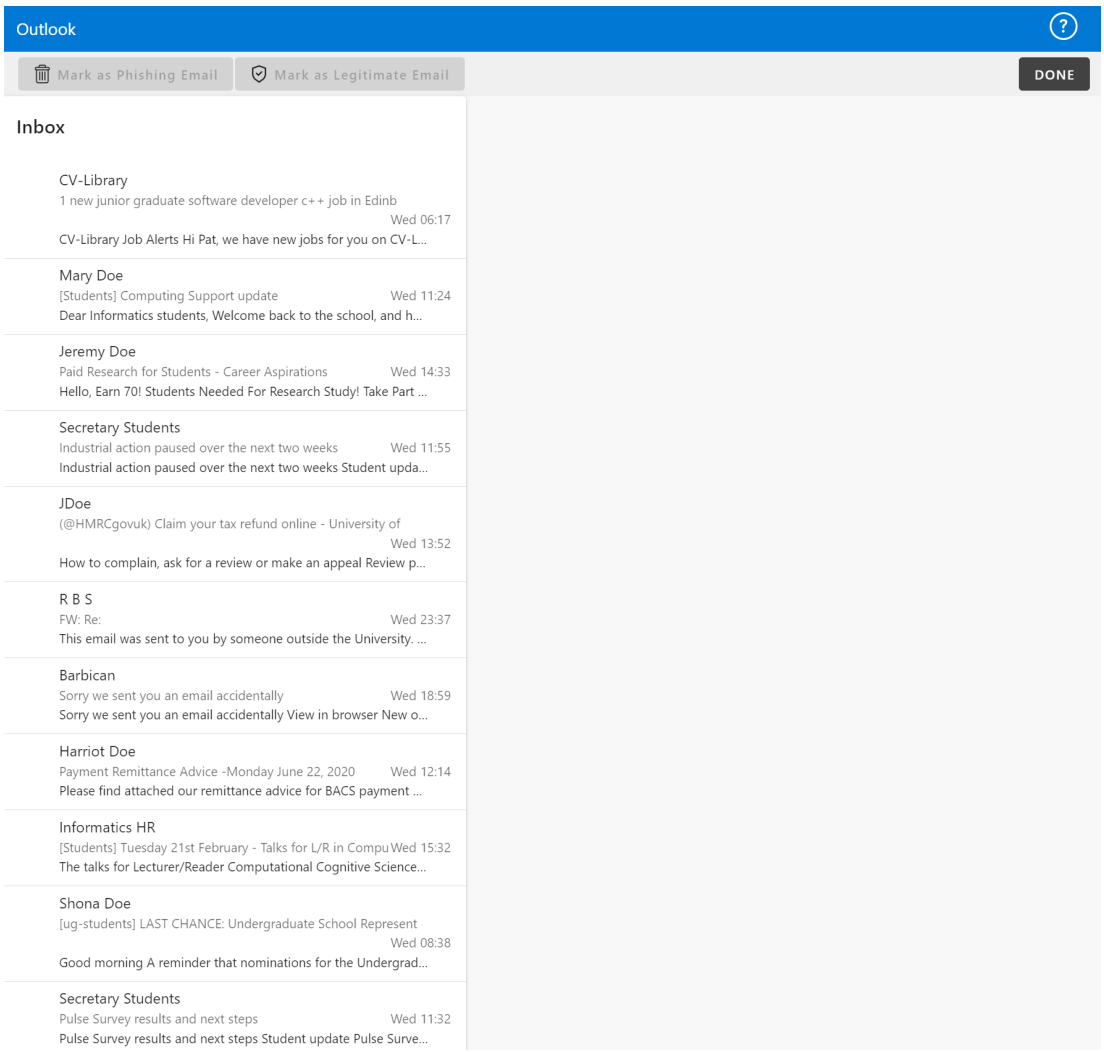


Figure B.11: Modified Task Page - Control Group User Flow.

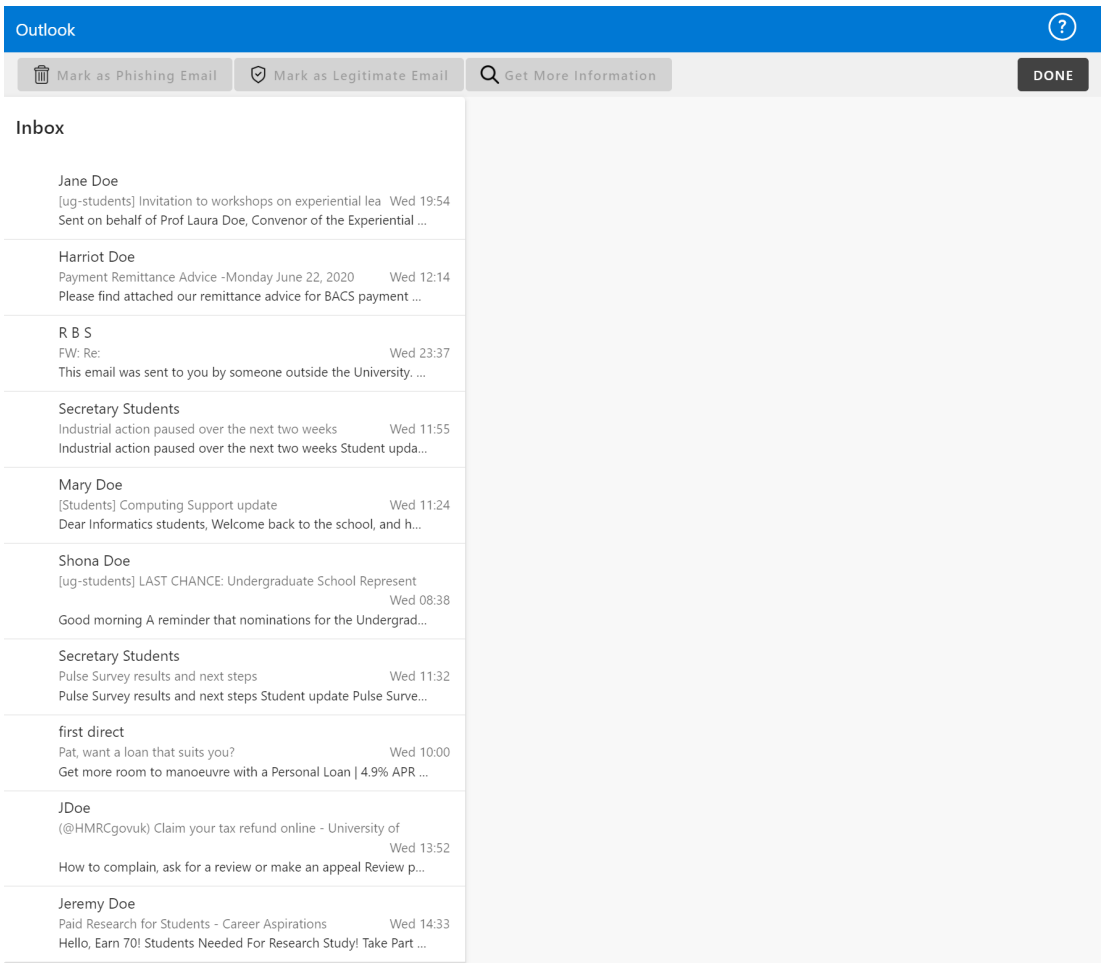


Figure B.12: Modified Task Page - Experimental Group User Flow.

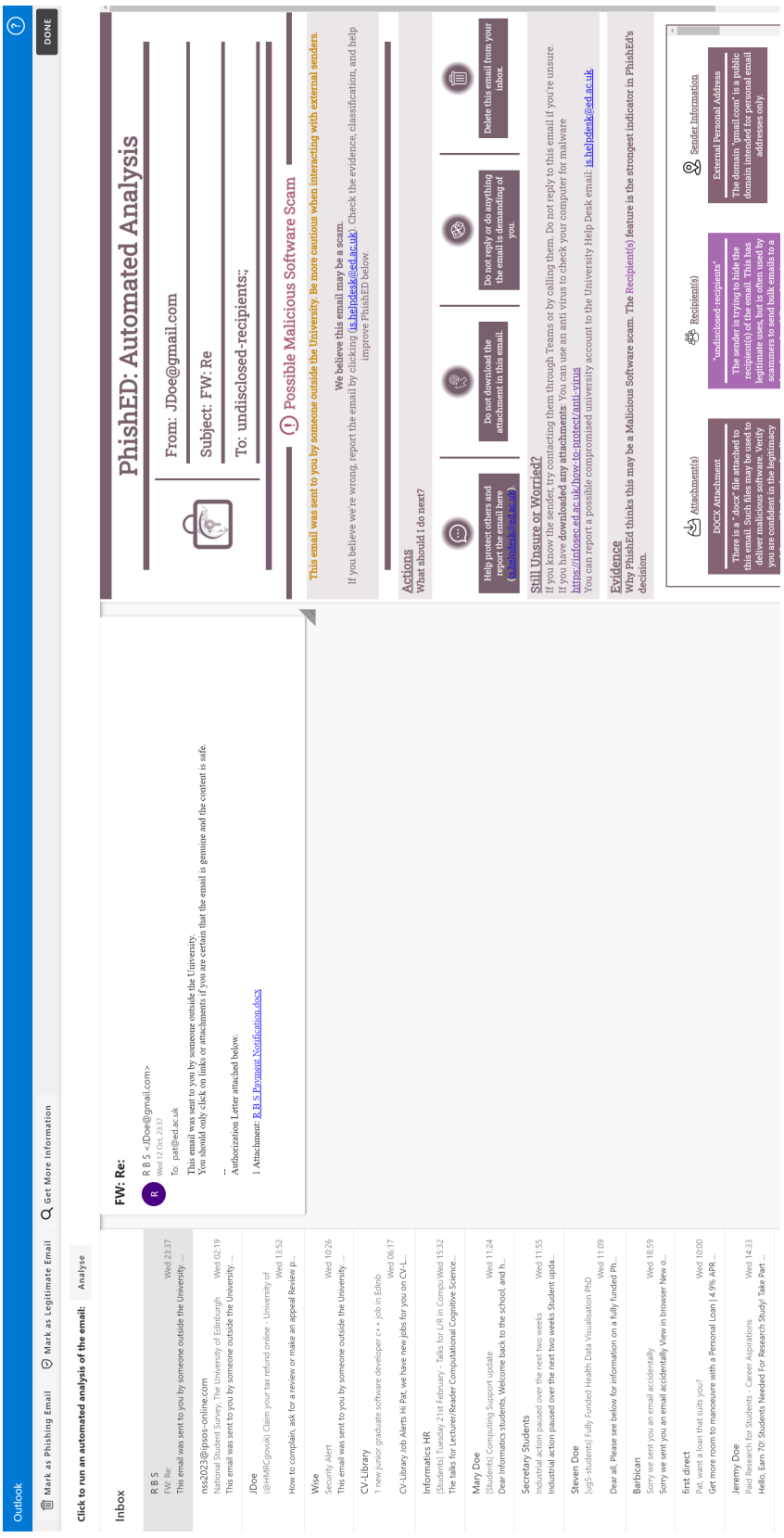
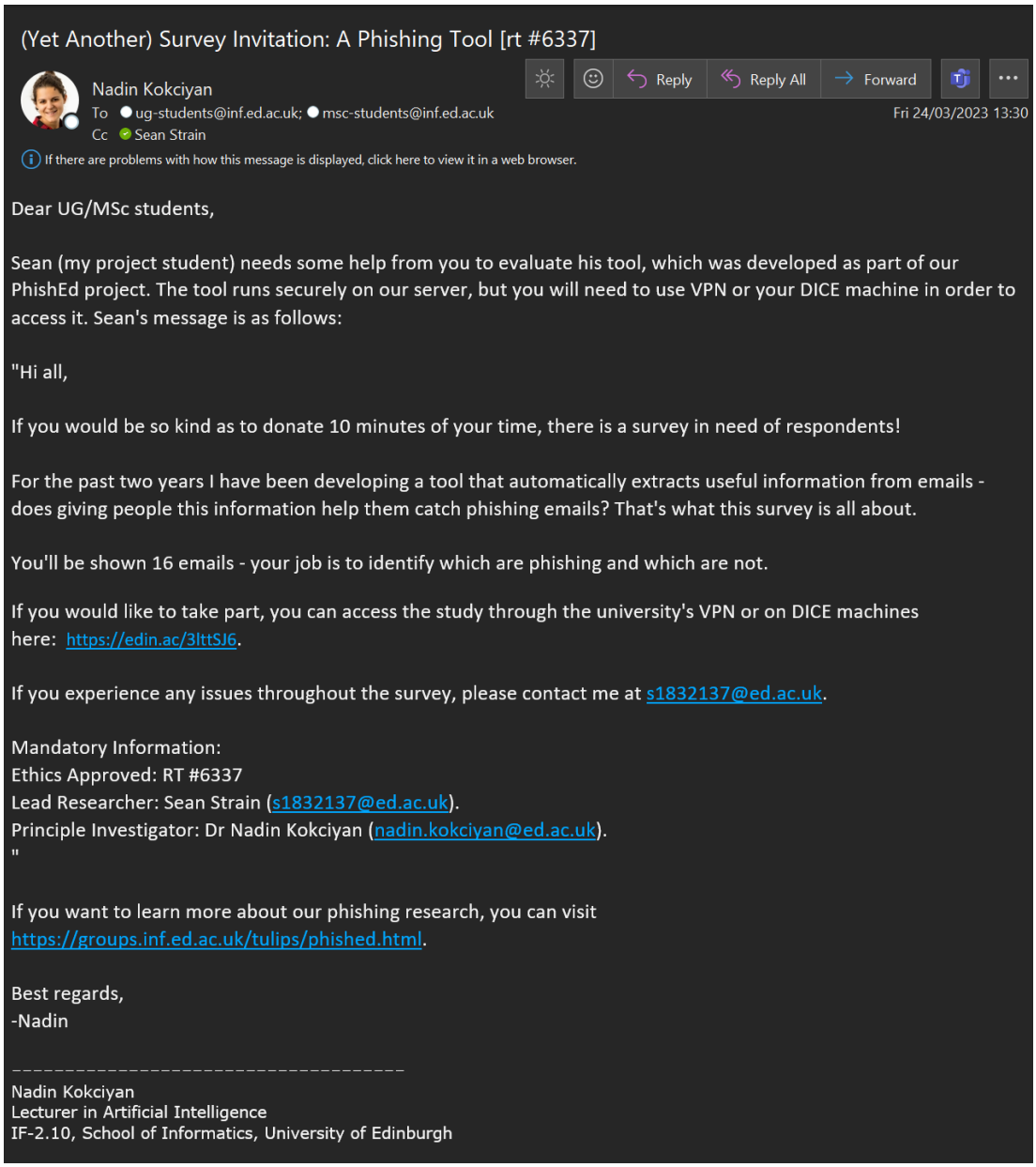


Figure B.13: Modified Task Page - Experimental Group User Flow. Analysis and Email displayed. Rotated $90°$ anticlockwise.

Appendix C

Participant Recruitment Email



(Yet Another) Survey Invitation: A Phishing Tool [rt #6337]

Nadin Kokciyan
To ug-students@inf.ed.ac.uk; msc-students@inf.ed.ac.uk
Cc Sean Strain

Reply | Reply All | Forward

Fri 24/03/2023 13:30

If there are problems with how this message is displayed, click here to view it in a web browser.

Dear UG/MSc students,

Sean (my project student) needs some help from you to evaluate his tool, which was developed as part of our PhishEd project. The tool runs securely on our server, but you will need to use VPN or your DICE machine in order to access it. Sean's message is as follows:

"Hi all,

If you would be so kind as to donate 10 minutes of your time, there is a survey in need of respondents!

For the past two years I have been developing a tool that automatically extracts useful information from emails - does giving people this information help them catch phishing emails? That's what this survey is all about.

You'll be shown 16 emails - your job is to identify which are phishing and which are not.

If you would like to take part, you can access the study through the university's VPN or on DICE machines here: https://edin.ac/3lttSJ6.

If you experience any issues throughout the survey, please contact me at s1832137@ed.ac.uk.

Mandatory Information:
Ethics Approved: RT #6337
Lead Researcher: Sean Strain (s1832137@ed.ac.uk).
Principle Investigator: Dr Nadin Kokciyan (nadin.kokciyan@ed.ac.uk).
"

If you want to learn more about our phishing research, you can visit https://groups.inf.ed.ac.uk/tulips/phished.html.

Best regards,
-Nadin

Nadin Kokciyan
Lecturer in Artificial Intelligence
IF-2.10, School of Informatics, University of Edinburgh

Figure C.1: Email sent to School of Informatics mailing list to recruit participants.

Appendix D

Post-task Survey

The post-task survey was 6 questions long. The survey was delivered using Microsoft Forms [81]. Questions 1-5 were mandatory. Question 1 asked participants to supply their output file. Question 2-4 were demographic questions. Question 6 was a free-form question. Figure D.1 and Figure D.2 displays this survey rendered in a web browser.

The survey can be accessed at the following URL: `https://forms.office.com/e/LKqyfH5LEu`.

Participants were permitted to give additional, free-form feedback from question 6 of the post-task survey. The question asked was as follows: "Do you have any other feedback for us, such as problems you had during the study or suggestions/advice?". 5 participants responded. Response 5 was merely the single character '-'. All responses are shown in Table D.1.

Response	Group	Feedback
1	Experimental	'I now trust nothing and no one.'
2	Control	'The e-mail viewer windows was a bit narrow'
3	Experimental	'The e-mail examples could include more phishing types (like people pretending to be family members) and more legitimate non-university organizations, or more phishing e-mails that look like promotion e-mails and vice versa.'
4	Control	'the suggestions are very clear, however its tricky to recognize the phishing emails.'
5	Experimental	'-'

Table D.1: Feedback received from participants from question 6 of the post-task survey.

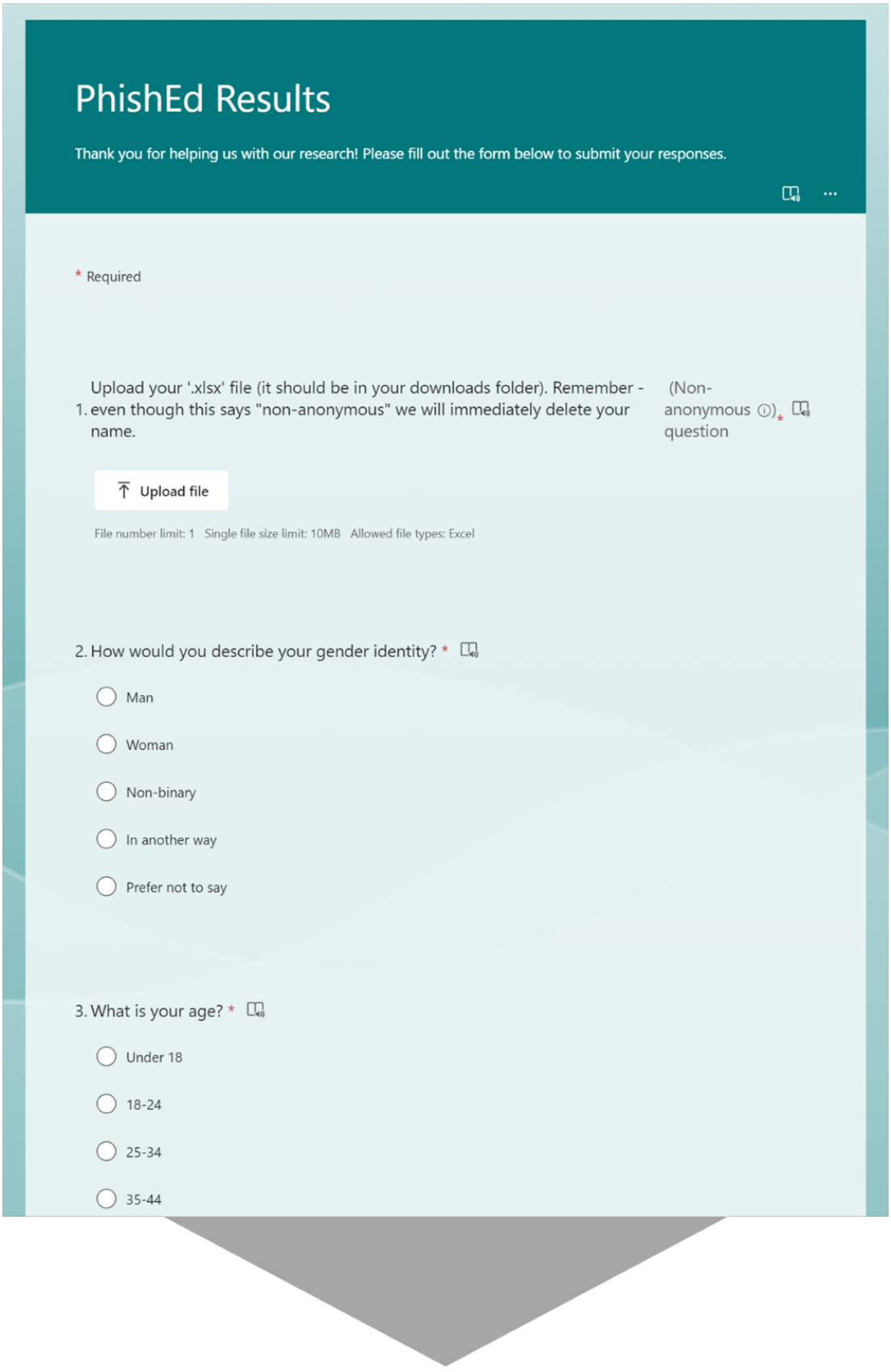


Figure D.1: Post-task Survey Part 1 of 2.

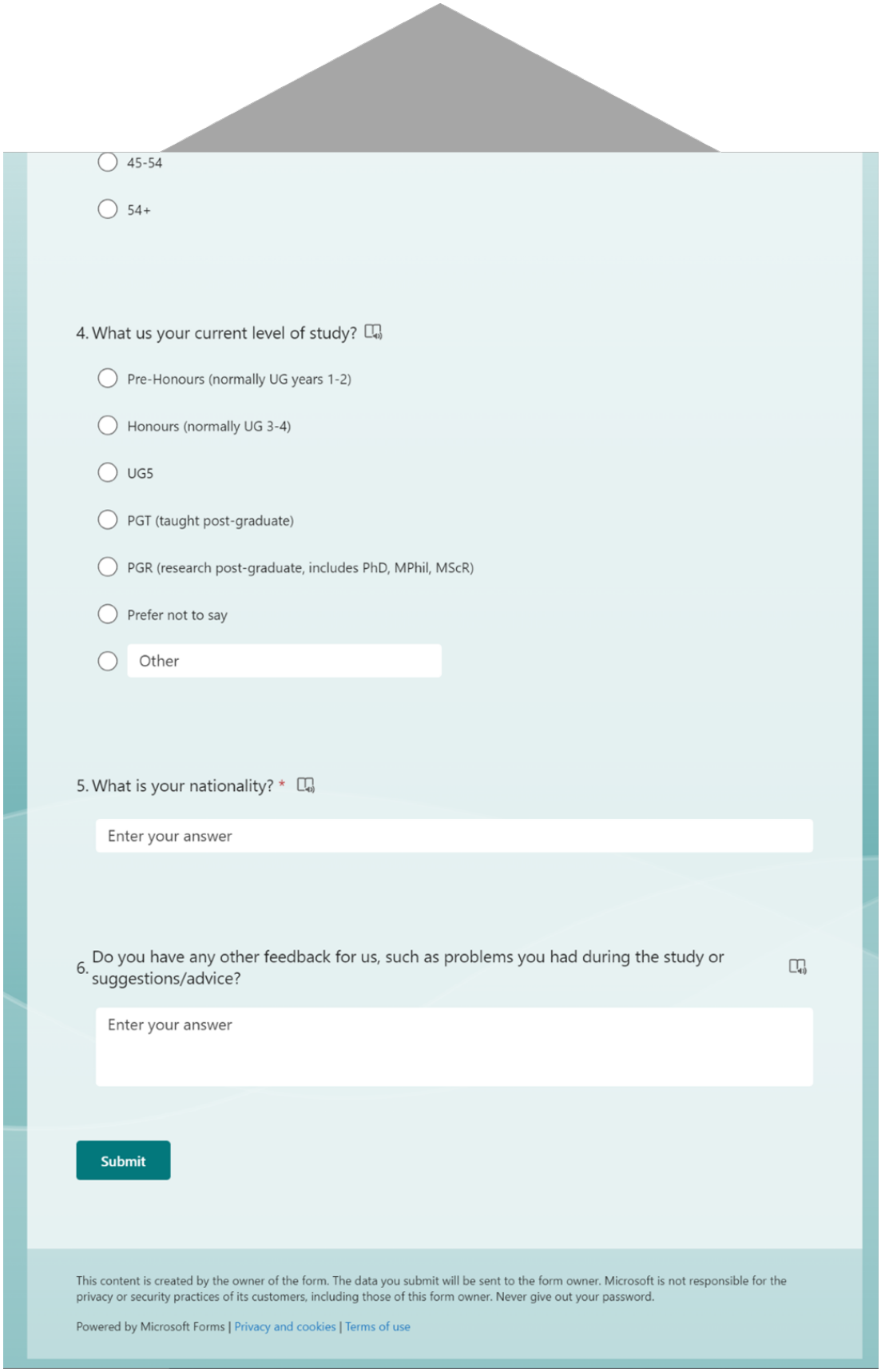


Figure D.2: Post-task Survey Part 2 of 2.

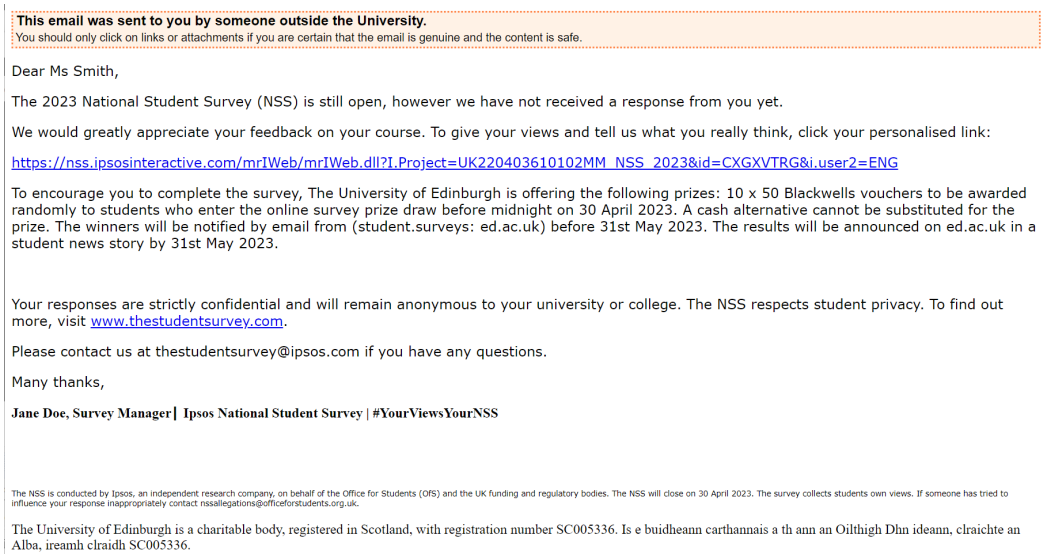
Appendix E

Materials Employed in User Study

This appendix evidences the emails and corresponding analyses of emails 4, 5, 6, and 14. These emails were chosen as they each were noted as particularly interesting in Chapter 6. All other emails are supplied in the supplementary materials submitted with this report.

Email	Subject	Truth Label	Classification
0	1 new junior graduate software developer c++ job in Edinburgh	Legitimate	Legitimate
1	FW: Re:	Phishing	Phishing
2	(@HMRCgovuk) Claim your tax refund online - University of Edinburgh -	Phishing	Phishing
3	Industrial action paused over the next two weeks	Legitimate	Legitimate
4	National Student Survey, The University of Edinburgh	Legitimate	Legitimate
5	Paid Research for Students - Career Aspirations	Phishing	Phishing
6	Payment Remittance Advice -Monday June 22, 2020	Phishing	Phishing
7	Pulse Survey results and next steps	Legitimate	Legitimate
8	Pat, want a loan that suits you?	Legitimate	Legitimate
9	Sorry we sent you an email accidentally	Legitimate	Legitimate
10	Security Alert	Phishing	Phishing
11	[Students] Computing Support update	Legitimate	Legitimate
12	[Students] Tuesday 21st February - Talks for L/R in Computational Cognitive Science	Legitimate	Legitimate
13	[ug-students] Invitation to workshops on experiential learning	Legitimate	Legitimate
14	[ug-students] LAST CHANCE: Undergraduate School Representative	Legitimate	**Phishing**
15	[ug5-students] Fully Funded Health Data Visualisation PhD Position in St. Andrews	Legitimate	Legitimate

Table E.1: Emails used in the study with subject, truth label, and analysis classification. If the analysis classification differs from the truth label it is shown in bold.



This email was sent to you by someone outside the University.
You should only click on links or attachments if you are certain that the email is genuine and the content is safe.

Dear Ms Smith,

The 2023 National Student Survey (NSS) is still open, however we have not received a response from you yet.

We would greatly appreciate your feedback on your course. To give your views and tell us what you really think, click your personalised link:

https://nss.ipsosinteractive.com/mrIWeb/mrIWeb.dll?I.Project=UK220403610102MM_NSS_2023&id=CXGXVTRG&i.user2=ENG

To encourage you to complete the survey, The University of Edinburgh is offering the following prizes: 10 x 50 Blackwells vouchers to be awarded randomly to students who enter the online survey prize draw before midnight on 30 April 2023. A cash alternative cannot be substituted for the prize. The winners will be notified by email from (student.surveys: ed.ac.uk) before 31st May 2023. The results will be announced on ed.ac.uk in a student news story by 31st May 2023.

Your responses are strictly confidential and will remain anonymous to your university or college. The NSS respects student privacy. To find out more, visit www.thestudentsurvey.com.

Please contact us at thestudentsurvey@ipsos.com if you have any questions.

Many thanks,

Jane Doe, Survey Manager| Ipsos National Student Survey | #YourViewsYourNSS

The NSS is conducted by Ipsos, an independent research company, on behalf of the Office for Students (OfS) and the UK funding and regulatory bodies. The NSS will close on 30 April 2023. The survey collects students own views. If someone has tried to influence your response inappropriately contact nssallegations@officeforstudents.org.uk.

The University of Edinburgh is a charitable body, registered in Scotland, with registration number SC005336. Is e buidheann carthannais a th ann an Oilthigh Dhn ideann, clraichte an Alba, ireamh clraidh SC005336.

Figure E.1: Email 4 - Truth Label: Legitimate.

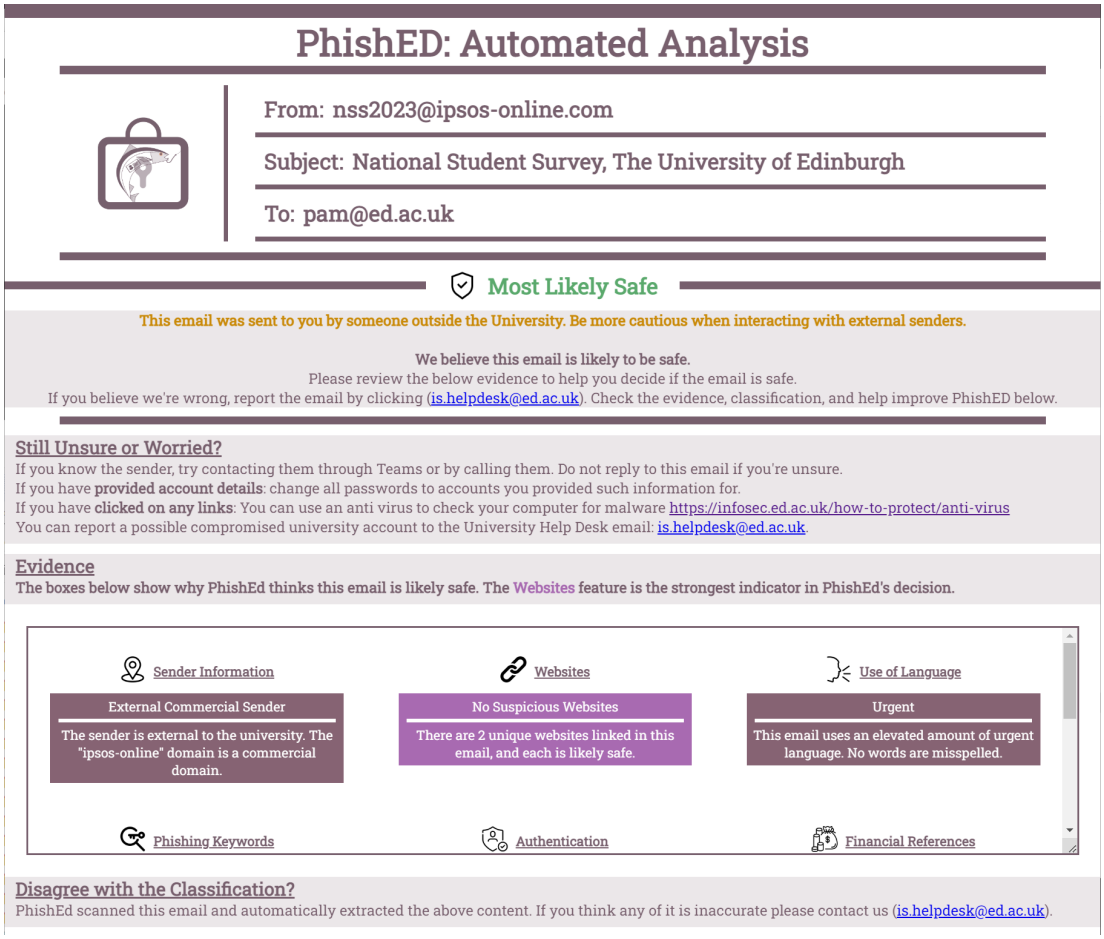


Figure E.2: Email 4 Analysis - Truth Label: Legitimate.

Hello,

Earn 70! Students Needed For Research Study! Take Part In A 60 Minute Online Research Interview About Career Aspirations and Goals! A British Research Company Needs Your Help!

What's Involved!

We would like to learn more about people's personal, career aspirations at different stages of their career. Taking part will involve speaking to our researcher, based in the UK via Zoom. This is a confidential research study, and we will NOT ask for any personal information. We are only interested in people's personal perspectives. The purpose of this study is to gain some perspective about what motivates people to work in different industries.

If you have any queries/concerns about the legitimacy of the project, My LinkedIn URL is in my email signature.
https://www.linkedin.com/in/laura-ivory-49569021/?originalSubdomain=uk

Payment We will pay you 70 for 60 minutes of your time. We are based in the UK, so we can pay you via PayPal/Amazon Voucher OR Bank transfer. We pay within 7 working days of your interview.

Date Wednesday 30th November, Friday 2nd December, Monday 5th December, Tuesday 6th December, Wednesday 7th December, or Friday 9th December

Time 5pm or 7pm (60 minutes)

*****Apply Now*****
If you would like to learn more about this project, please **CLICK HERE** to apply now.

Best regards,

Jeremy Doe
Fieldwork Manager
Recruitment for Research
tel: 0207 --- ----
mobile: 07---------
email: jeremydoe@recruitmentforresearch.com
web: www.recruitmentforresearch.com
https://www.linkedin.com/in/laura-ivory-49569021/?originalSubdomain=uk

Recruitment for Research

20-22 Wenlock Road London N1 7GU

Click here to Unsubscribe

Figure E.3: Email 5 - Truth Label: Phishing.

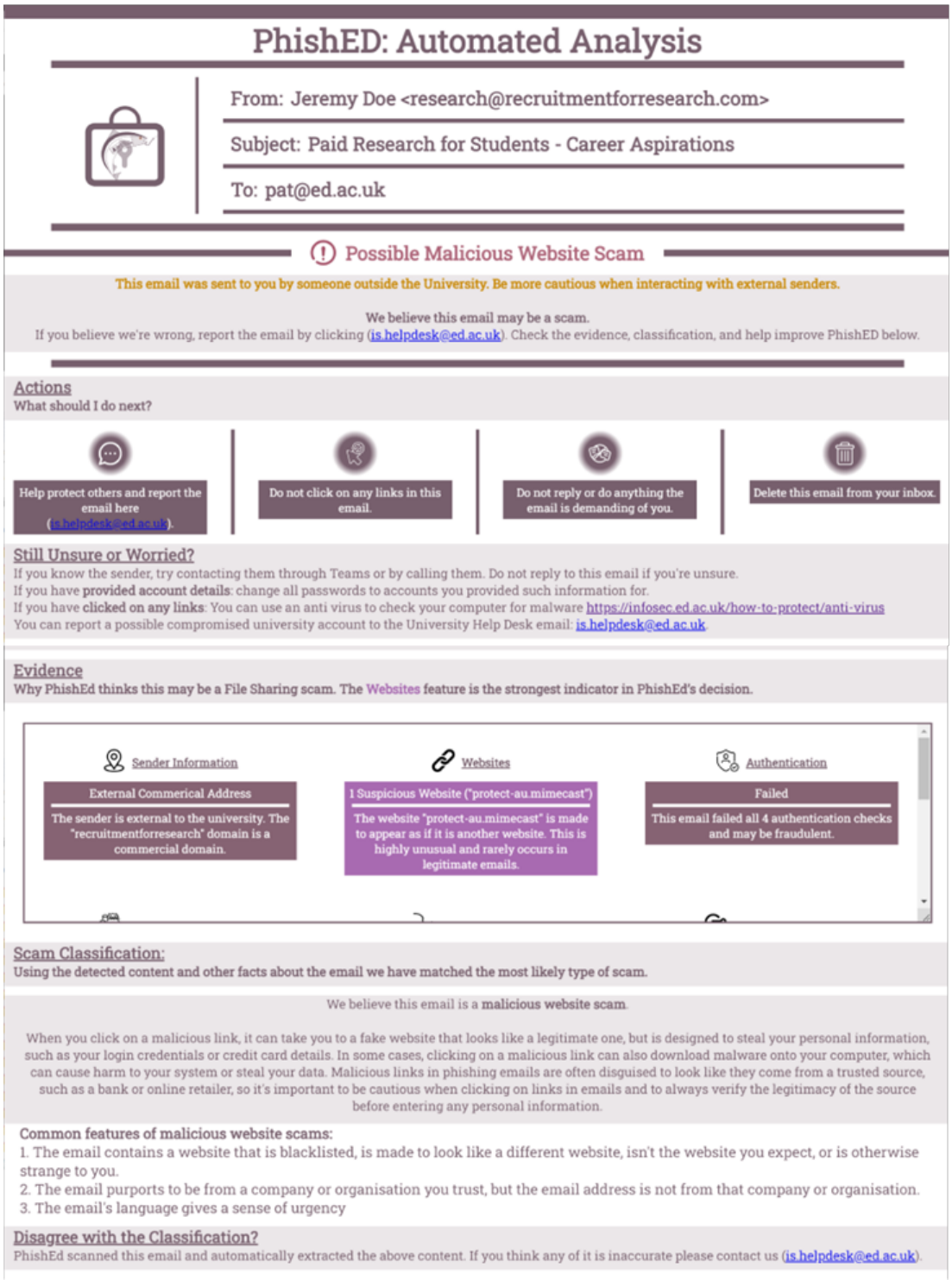
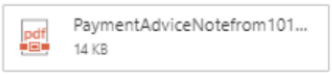


Figure E.4: Email 5 Analysis - Truth Label: Phishing.



Please find attached our remittance advice for BACS payment made.

This payment should reach your bank account within 3 working days of remittance date shown.

For partnership agreement invoices, please continue to email as per agreement.

Kind Regards,
Accounts Payable

Figure E.5: Email 6 - Truth Label: Phishing.

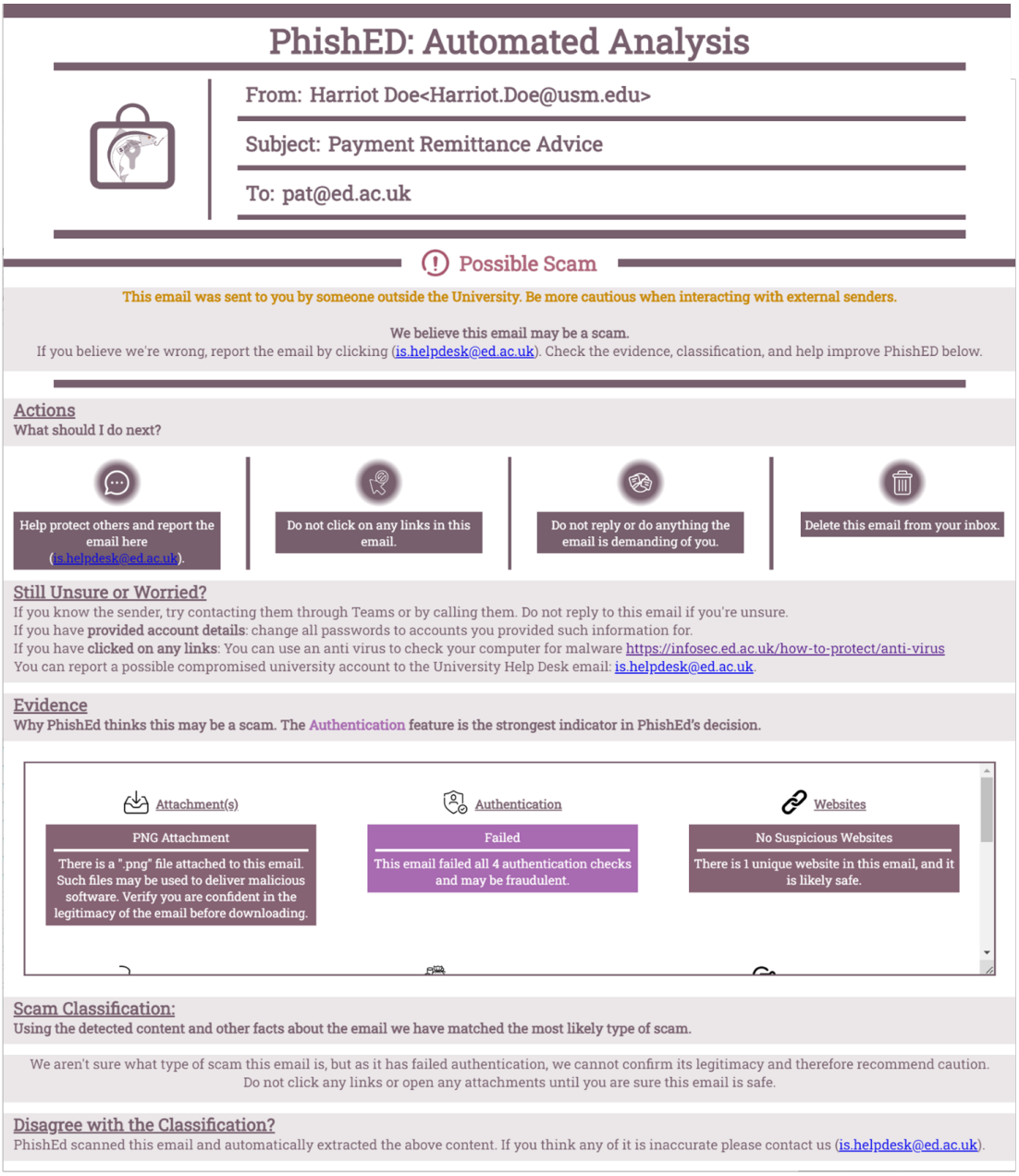


Figure E.6: Email 6 Analysis - Truth Label: Phishing.

Good morning

A reminder that **nominations for the Undergraduate School Representative role close at 12.00pm TODAY.**

Please consider nominating yourself to ensure that Informatics students are properly represented next academic year.

Best wishes

Shona

From: Shona Doe
Sent: 14 February 2023 09:00
To: ug-students@inf.ed.ac.uk
Subject: IMPORTANT: Undergraduate School Representative
Importance: High

Dear UG Students

I have sent you information about the Students Association Undergraduate School Representative role previously, but am contacting you again now because **there is currently no-one from Informatics standing for the role.**

Nominations close at 12:00pm TOMORROW (Wednesday 15 February), and Informatics and the Students Association are keen to ensure that the role isnt vacant and that the Schools students are properly represented in the coming academic year.

Please consider nominating yourself to ensure that students like you have a voice next year. School Representative roles are designed to fit flexibly around your studies and other commitments, and are fully supported by staff within the School and the Students' Association. You could help to enhance the student experience in Informatics, whilst also gaining valuable transferable skills and connecting with students and staff from across the School and beyond.

Find out more about the role and submit your nomination here: www.eusa.ed.ac.uk/elections

Shona Doe
School of Informatics
The University of Edinburgh
Informatics Forum
10 Crichton Street
Edinburgh
EH8 9AB

Shona.Doe@ed.ac.uk
Can also be contacted via Microsoft Teams

School of Informatics | The University of Edinburgh

Figure E.7: Email 14 - Truth Label: Legitimate.

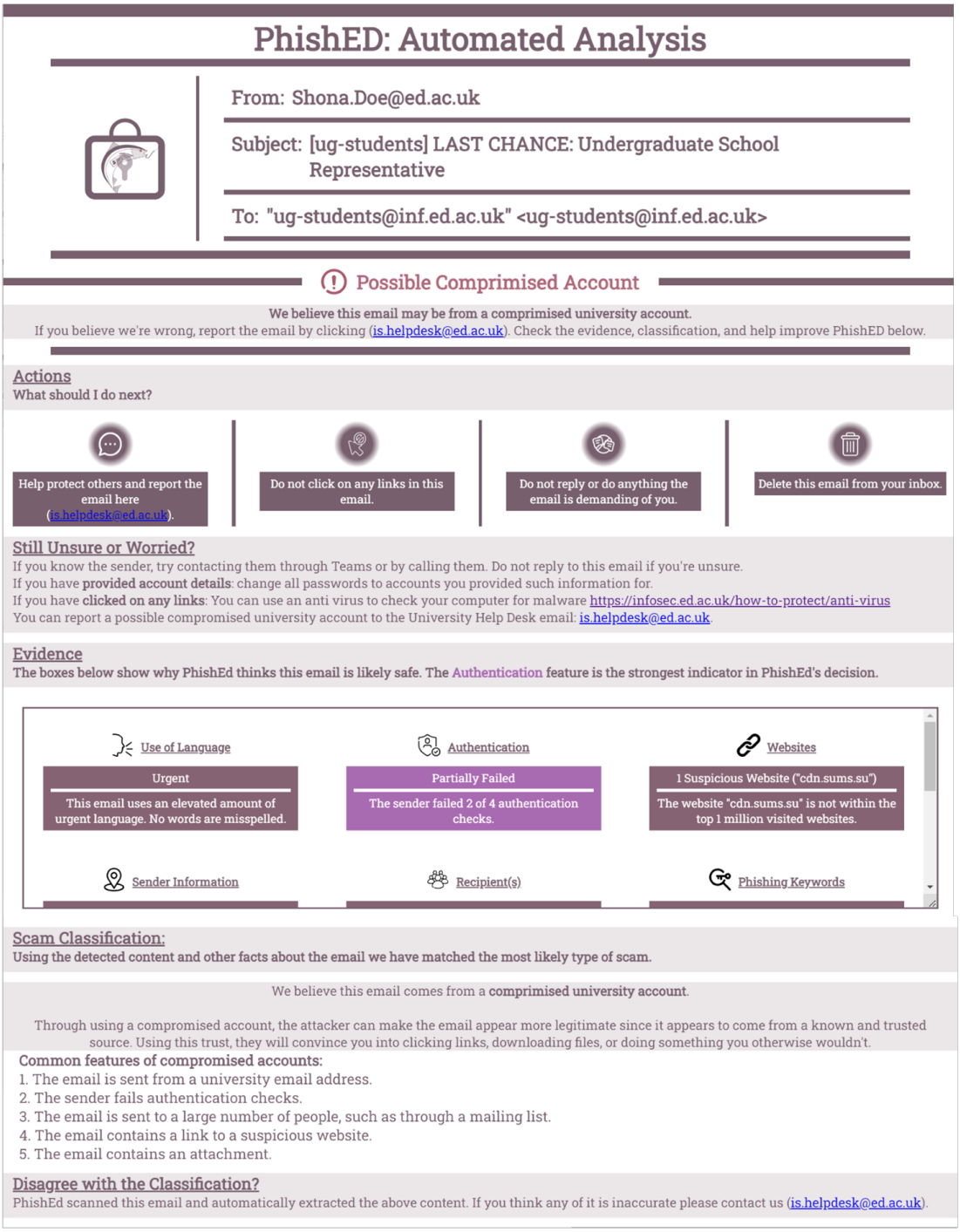


Figure E.8: Email 14 Analysis - Truth Label: Legitimate. Note that the overall classification is wrong in this instance.

Appendix F

Microsoft Outlook Add-in

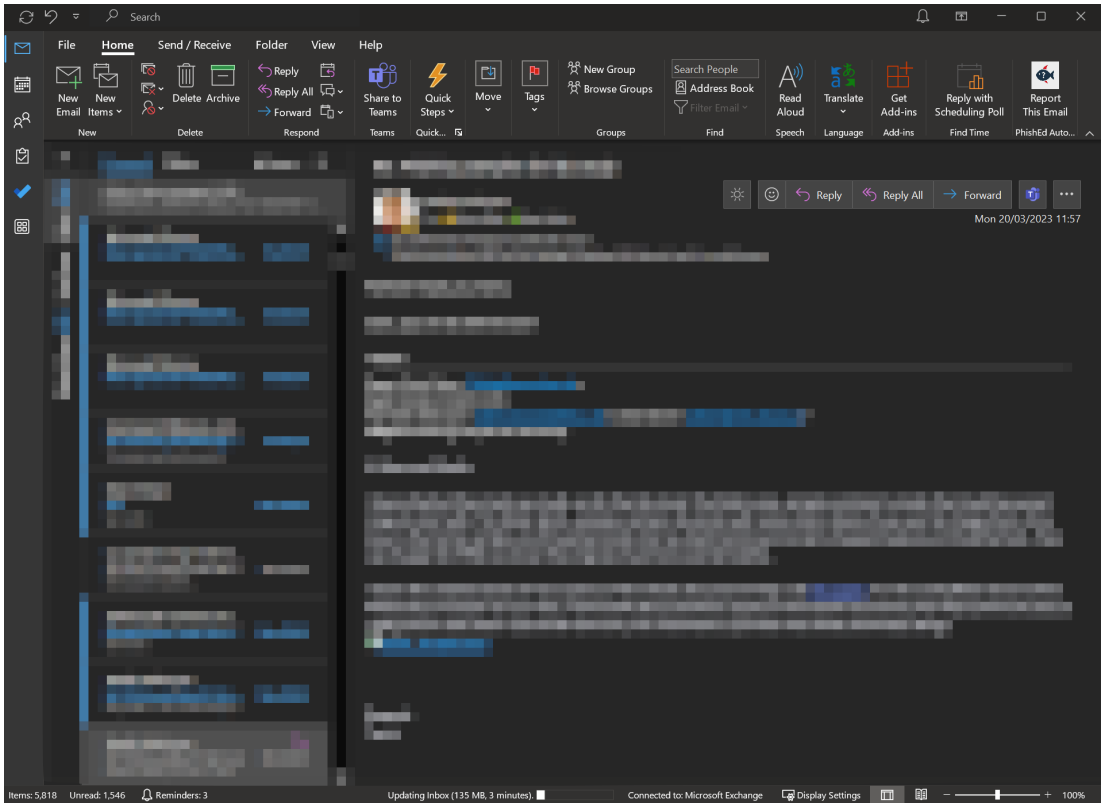


Figure F.1: Microsoft Outlook Desktop Application. PhishEd report button shown in top-right. Email pixelated for the protection of the sender's privacy.

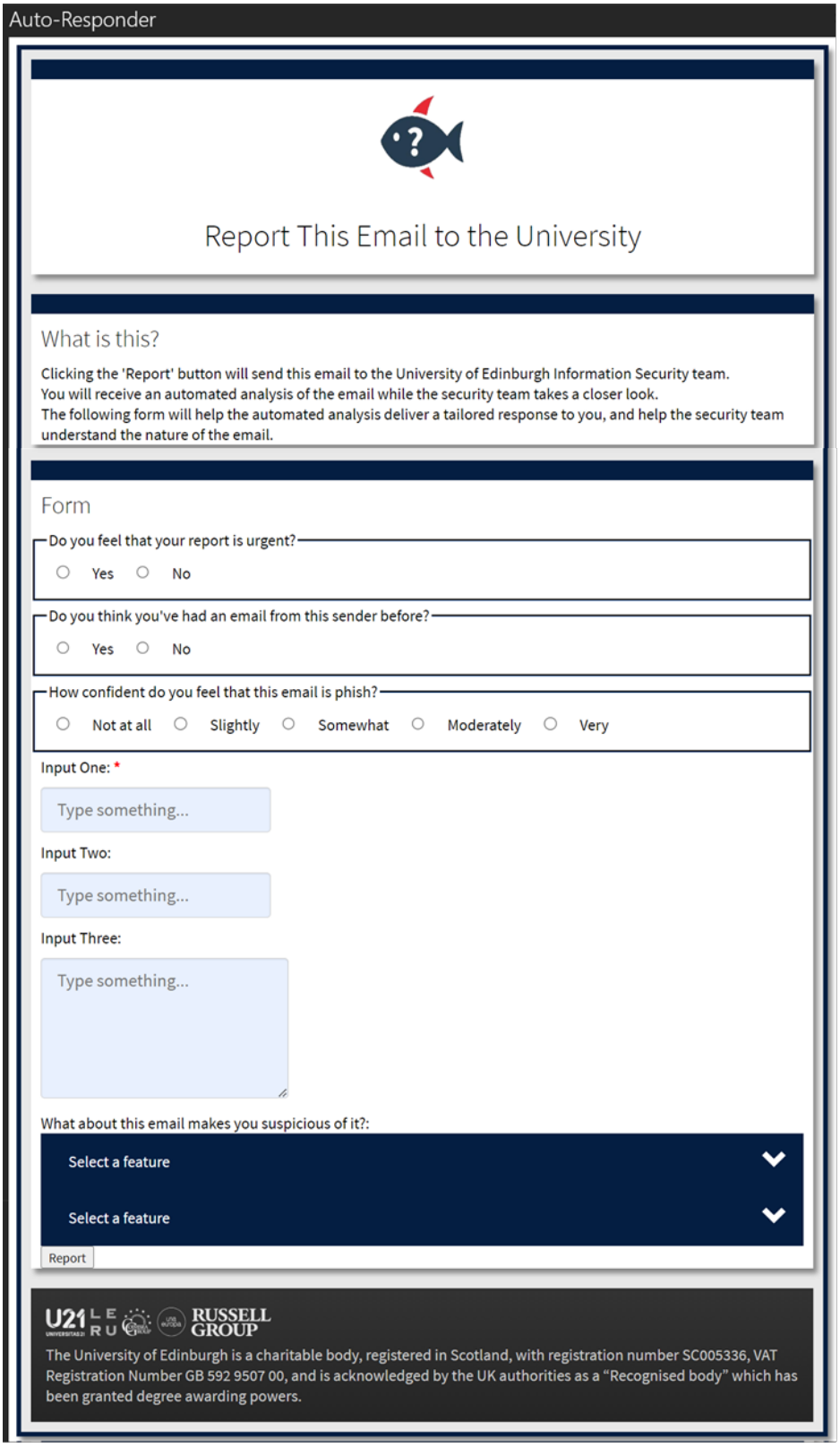


Figure F.2: Microsoft Outlook Desktop Application. Sidebar shown once PhishEd report button has been clicked.

Appendix G

Generative Language Models in Phishing

Generative language models that create convincing human-like text have become increasingly prevalent in the last two years and are now freely available to use [34]. Brown et al. specifically highlights the potential for their model - GPT-3 - to be misused to generate phishing/spam emails [12]. These models would allow an attacker to automate their phishing campaigns by generating high-quality phishing emails that no longer possess typical phishing characteristics, such as misspellings. It would be negligent to ignore the potential effects such models may have on the email analysis module and its applications - and, indeed, on the field of phishing as a whole. Given the recency of these models, research on this topic is limited. Lohn and Jackson speculate as to how models such as these may affect the rates of phishing campaign detection and user susceptibility [68], suggesting it may increase clickthrough rates tenfold and invalidate decades of anti-phishing training. However, at the time of writing, phishing emails created by these models have not been quantitatively compared to human-written emails in any context. Therefore, the consequences on the field of phishing can only be speculated. Nonetheless, it is important to keep such models in mind in future, should their usage become prevalent. In which case, it may be possible for the email analysis module to incorporate AI generated text detection software in its analysis.