Design and Motivational Factors for an Effective Educational Game for Malicious Email Detection

by

Vignesh Kumar Karuppasamy

A thesis
submitted to the University of Ottawa
in partial fulfillment of the
requirements for the degree of
Master of Computer Science (MCS)

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Examining Committee

Supervisor	Prof. David Knox School of Engineering Design and Teaching Innovation School of Electrical Engineering and Computer Science University of Ottawa
Internal Examiner	Prof. Hussein Al Osman School of Electrical Engineering and Computer Science University of Ottawa
External Examiner	Prof. AbdelRahman Abdou School of Computer Science Carleton University

Declaration of Authorship

I hereby certify that this thesis is entirely my own original work except where otherwise indicated. I am aware of the University's regulations concerning plagiarism, including those concerning consequent disciplinary actions. Any use of the works of any other author, in any form, is properly acknowledged at their point of use.

Abstract

Phishing email attacks remain one of the dominant and ever-evolving cybersecurity threats facing organisations, demanding defensive measures to adapt continually. Existing anti-phishing education methods have narrowly focused on certain aspects of phishing detection, such as URL verification, instead of incorporating insights from learning theory, such as the role of motivation and cognitive message processing in phishing detection. In particular, the effectiveness of collaborative game elements in anti-phishing training remains largely unstudied. This thesis addresses these gaps through two interconnected studies. First, we analyse the Motivated Strategies for Learning Questionnaire (MSLQ) across different instruction modes, including remote, in an undergraduate engineering design course. From this, we derive a validated motivational construct, which we then apply to the anti-phishing training domain. Second, we design and implement *Phooled!*, a scalable collaborative anti-phishing training game, and evaluate it through a comparative study. We find that collaborative training improves phishing detection efficiency, measured as a balance between accuracy and speed, by 32% compared to the control group. Additionally, we demonstrate that motivation predicts phishing detection improvement, and identify key determinants of phishing email detection difficulty.

Acknowledgements

I am sincerely grateful to my supervisor, Prof. David Knox, for his guidance, feedback, and support over the years. I would also like to thank Prof. Hussein Al Osman and Prof. AbdelRahman Abdou for their time in reviewing this thesis.

I am grateful to the members of the Phishducation research group who assisted with the email labelling and consensus process. I also extend my thanks to all the participants who took part in the *Phooled!* study, in particular the students from the Faculty of Engineering Secondary School (FESS) and the staff from the uOttawa Cybersecurity Office.

Finally, I thank my family, C. Karuppasamy, Uma Maheswari, and Kishore Kumar, for their endless support and encouragement throughout my studies.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANCOVA Analysis of Covariance 81

ANOVA Analysis of Variance 25

CI Confidence Interval 31

EMM Estimated Marginal Mean 84

FNR False Negative Rate 90

FPR False Positive Rate 90

FSM Finite State Machine 71

GLM Generalised Linear Model 92

GLMM Generalised Linear Mixed Model 81

HSM Heuristic-Systematic Model 11

HyFlex Hybrid Flexible 20

IES Inverse Efficiency Score 12

LMM Linear Mixed Model 81

MPA Motivation and Perceived Ability 4

MSLQ Motivated Strategies for Learning Questionnaire 2

RQ Research Question 19

SCAM Suspicion, Cognition, and Automaticity Model 11

SN Subjective Norms 96

WS WebSocket 69

Chapter 1

Introduction

Phishing has been among the most prevalent and costly cybersecurity threats since the earliest days of the internet. Despite advances in machine learning methods used by the automated spam filters of email providers, a small percentage of phishing emails continue to reach the inboxes of users. This is primarily because phishing is an adaptive threat and malicious attackers refine their phishing methods over time to evade detection. Therefore, to avoid such attacks, organisations need to arm their email users, the final line of defence against phishing, with skills to detect phishing attacks and avoid falling victim to them.

The anti-phishing training domain includes various commercial products, and research in this area has steadily grown to address the cost of such attacks borne by organisations. There are two dominant approaches: embedded training and game-based training. Embedded training is the more traditional approach, where an organisation simulates a phishing attack to identify users for remedial training measures such as online courses. On the other hand, gamification seeks to engage users in an educational game to teach skills that translate to the real-world phishing detection task. While both approaches have shown promise, significant gaps have been identified in game-based training, particularly for the learning material and user interaction mechanisms.

Existing anti-phishing training games focus narrowly on certain aspects of phishing detection, such as verifying URLs in the email, rather than using a more holistic approach,

involving the entire email body and email metadata. Similarly, while collaborative learning has been established as an effective means of education in other domains, the study of its applicability in digital anti-phishing games has been limited. Finally, the interplay between individual learner characteristics, such as motivation and cognitive processes of message parsing as drivers of detection performance improvement has not been studied extensively.

This thesis addresses these gaps through two interconnected studies. First, a reduced motivational construct is derived from the Motivated Strategies for Learning Questionnaire (MSLQ) survey administrated across different instruction modes in the context of several consecutive terms of a university engineering course. The second study designs, implements, and evaluates *Phooled!*, an online collaborative anti-phishing training game. Using *Phooled!*, we examine the effectiveness of remote collaborative education in the anti-phishing domain. Additionally, we measure the effectiveness of our approach by analysing the rich user-game interaction data generated through user interaction with our designed in-game interaction mechanics.

1.1 Problem Statement

Organisations, such as universities and private enterprises, increasingly rely on their email users to combat phishing threats against the organisation. Because automated mechanisms do not detect all anti-phishing attempts, they must train their email users to detect and evade phishing emails that reach their inbox. However, existing anti-phishing training approaches have not worked well. Anti-phishing training games in particular have narrow user-game interaction mechanisms that do not incorporate insights from related domains such as cognitive processing, and motivational aspects of behavioural psychology have been largely ignored. These are important if learners are to transfer their learning to real-world email environments. Traditional approaches, such as in-person courses or board games, require an expert in the subject matter to impart learning, which is a constraint given the shortage of security personnel and instructors. There is significant scope for scaling collaborative education through digital games, playable remotely, for widespread

usage in organisational settings. In this work, we analyse the effectiveness of remote and collaborative anti-phishing educational games.

1.2 Research Objectives

These are the research objectives for the thesis work:

1. To analyse the MSLQ survey across various instruction modes in a university organisational setting, and to derive a reduced motivational construct suitable for use in an anti-phishing education domain.
2. To design and implement *Phooled!*, a collaborative and remote anti-phishing training game that incorporates real-world phishing content and in-game interaction mechanics such as time pressure, role-play, and interactive feedback.
3. To evaluate whether collaborative digital anti-phishing games improve phishing detection after learning compared to single-player and/or control treatments.
4. To examine whether individual learner characteristics, specifically motivation and subjective norms, predict learning improvement, after or during training with *Phooled!*.
5. To investigate which elements of a phishing email undermine user phishing detection performance, through user-game interaction log data.

These broad objectives are refined into specific research questions in Chapter 2, grounded in the gaps identified in the literature review.

1.3 Contributions

This thesis makes the following contributions to anti-phishing education research:

1. **Reduced motivational construct.** By analysing the MSLQ survey across different instruction modes in a university course, we establish a dominant motivational construct that is then applied to the anti-phishing training domain.

2. **Design and implementation of Phooled!** We design and implement Phooled!, which we believe to be the first digital anti-phishing training game designed for remote and collaborative play, by incorporating holistic phishing email content alongside rich user-game interaction mechanics such as time pressure, role-play, and interactive feedback.

3. **Evaluation of collaborative training.** We evaluate the effectiveness of remote collaborative anti-phishing training through a comparison of a single-player and multi-player version of *Phooled!*. We find that collaborative training improves phishing detection efficiency by 32%, compared with the control group.

4. **Role of individual characteristics.** We demonstrate that higher Motivation and Perceived Ability (MPA) scores predict greater phishing detection improvement, and that collaborative gameplay eliminates the opposing effects of peer and authority influence of subjective norms on learning.

5. **Evaluation of phishing detection difficulty.** Safe emails are shown to be harder for users to classify compared with phishing emails. Additionally, the only phishing cue that individually affects detection is the sender's email domain, while the aggregated presence of multiple cues determines overall sample difficulty. Moreover, we establish that users are generally overconfident in their phishing detection ability.

1.4 Publications

The study presented in Chapter 3 is published in the proceedings of the 2025 ASEE Annual Conference & Exposition:

- D. A. Knox and V. K. Karuppasamy. 2025. Motivation and Learning Strategies - What can university engineering design courses do to help students and what must students do? *2025 ASEE Annual Conference & Exposition*. ASEE Conferences [24].

1.5 Thesis Outline

The remainder of this thesis is organised as follows:

Chapter 2 surveys the literature on the existing anti-phishing training approaches, persuasive message processing theory, learning theory and the importance of motivation, and analyses the individual characteristics of user susceptibility to phishing content. As already stated, the chapter concludes with the specific and detailed research questions for the thesis.

Chapter 3 presents a study of an MSLQ survey across cohorts of an undergraduate engineering design course under three instruction modes: in-person, hybrid, and remote. Through exploratory factor analysis, we derive the MPA construct that is used to measure motivation in the anti-phishing study.

Chapter 4 describes the design and implementation of *Phooled!*, a collaborative anti-phishing training game. The chapter describes the game design, experiment design, and also the system design choices that are implemented in *Phooled!*.

Chapter 5 presents the experimental results that answer each research question. It examines: phishing detection improvement across treatment conditions, the role of individual characteristics in predicting phishing learning, the evaluation of in-game interaction mechanics, and the determinants of phishing email detection difficulty. The chapter concludes with an analysis of participant feedback from a feedback survey and from focus groups.

Chapter 6 summarises the findings from the *Phooled!* study and discusses the limitations of our experiments, suggesting productive directions for future work.

Chapter 2

Literature Review

Phishing is an email-based cybersecurity attack used by malicious actors to harm unsuspecting users. Such attacks typically take one of two forms: deceiving users into divulging confidential information on a fraudulent website, or installing malware to their computing device (e.g., personal computers or smartphones) that relays confidential information back to the attacker. The threat surface for such attacks grows in organisations, such as universities or private companies, where each email user represents a potential vulnerability. The annual cost of a successful phishing attack on an organisation averaged $14.8 million in 2021 [65], and by 2023, 71% of organisations experienced at least one such successful attack [66].

The primary line of defence against such phishing threats is automated spam filters from email providers such as Gmail or Outlook, either on the cloud or on the email client itself. These automated systems use machine learning models to classify and filter phishing emails before they reach a user's inbox. However, even state-of-the-art deep learning models are not perfect, achieving between 95–99% accuracy on phishing email classification [8, 80]. Consequently, a small percentage of phishing emails will always bypass installed automated systems. Moreover, phishing is an adaptive threat, where attackers continually evolve their strategies to evade existing detection systems. This was evident during the COVID-19 pandemic, when malicious actors used fear surrounding the outbreak as a core part of

their phishing strategy [4, 5].

Therefore, it is imperative that organisations arm their final line of defence, the email users themselves, through awareness campaigns and anti-phishing training against current and future adaptive threats. Anti-phishing training research has matured in the past two decades, particularly from 2007 onwards, when the initial literature in this domain was developed [51]. This literature review is organised as follows: we look at the foundations and approaches of existing anti-phishing training (Sections 2.1.1, 2.1.2, 2.1.3), then examine the theoretical foundations of persuasive message processing and learning (Sections 2.2, 2.3), followed by motivation and behavioral psychology (Section 2.4), determinants of phishing susceptibility (Sections 2.5, 2.6), and conclude with the specific research questions (Section 2.7) derived from the objectives stated in Chapter 1.

2.1 Anti-Phishing Training

Embedded training has been the primary means of anti-phishing training in the existing literature [43, 51]. In embedded training, an organisation sends a simulated phishing email attack across their network. Users who are deceived by this simulated attack (e.g., by clicking on an external link in the email) are signed up for anti-phishing training. The remedial measures might be simple, such as an infographic promoting phishing email detection tips, or more complex, such as in-person training.

More recently, researchers have focused on the development of games to educate users on phishing detection, often referred to as anti-phishing learning games [69] or serious games [20]. This gamified approach seeks to introduce engaging learning material that would be attractive to learners compared with traditional methods. These games contain rich in-game interaction mechanics, like reward incentives and role-play, to promote user engagement and enable more immersive learning. We begin by first describing the traditional embedded training approach in Section 2.1.1. Next, we examine the current insights in the anti-phishing learning games in Section 2.1.2. Finally, we conclude with the specific in-game user interaction mechanisms incorporated by these games and the gaps identified

in these existing games in Section 2.1.3.

2.1.1 Embedded Anti-Phishing Training

Kumaraguru et al. [50] used a simulated phishing attack that directed users to an external website where their username and password were phished. The phished users were shown learning material such as a comic strip with anti-phishing tips. Embedded training was found to be more effective than non-embedded treatment where the learning material is sent directly by email [51]. Most users who clicked on the phishing link did so within 8 hours [50], reinforcing the notion that automated filters and manual intervention by an organisation's cybersecurity office would not be effective. Users who saw this training material twice performed better than those who saw it only once [50]. Additionally, the false positive rate, i.e. the rate of legitimate emails incorrectly classified as phishing, did not increase after training [50].

88% of users who clicked on the link also divulged their credentials on the external website [52]. This validates most studies' approach of training users to identify phishing, before engaging with any actions required by the email. No difference was found in phishing detection skills between groups that received generic training versus spear phishing-specific training [52]. McElwee et al. [55] found that organisational incentives and rewards did not predict phishing avoidance, but repeated training did.

A more immersive version of embedded training was proposed by Alnajim and Munro [6] where a Domain Name System (DNS) re-routing mechanism was introduced between the user's email client and the internet. This mechanism intercepted and rerouted network requests from malicious email links to show training material instead. Advocates of such methods claim that an embedded mechanism, where the act of *getting phished* occurs in the user's real-world email reading environment (e.g., Microsoft Outlook) using their inherent email processing strategies, as the best setting for such interventions, referred to as an authentic learning environment [38].

More recent research, however, has questioned the effectiveness of embedded training. Ho et al. [40] analysed embedded training effects across a large organisation and found

that trained employees were only 1.7% less likely to fall victim to a simulated phishing attack than untrained employees. The study also found that between 37% and 51% of users showed no engagement at all [40].

2.1.2 Anti-Phishing Learning Games

The seminal work in the gamified approach is *Anti-Phishing Phil* [72], a game developed to teach users to identify phishing URLs. Anti-phishing games have been shown to increase learner engagement and improve phishing detection skills [20, 75], improving overall phishing classification accuracy from around 68% to 86% ($n = 36$ participants) [61], with no increase in false positives.

Sheng et al. [72] found that the only effective treatment in their study was a game which reduced false positives, while traditional intervention material made the users more cautious, thereby decreasing accuracy. Other studies have connected these lab results to organisational settings on a larger scale. CJ et al. [20] showed that in an organisation (n = 8,071 participants), users' URL classification accuracy improved from 71% to 79% after playing a game focused on identifying phishing URLs. The next evolution came through shifting focus from URL classification to email sample contents. In *What.Hack*, the focus was on email classification rather than URLs [90]. *What.Hack* improved classification accuracy from 65% to 89%, while neither URL-based games nor traditional learning material showed any improvement [90].

A survey of 40 anti-phishing learning games found that certain key areas are largely being ignored by current literature [69]. In particular, the majority of games focus on URLs rather than on email content as a whole [69, 79]. Furthermore, many games focus on quizzes about concepts without requiring players to apply the knowledge being taught [69]. Even when games do teach specific procedures, they are game-specific actions like eating worms in a virtual pond, and not applicable to real-world phishing detection [69]. In most cases, existing in-game mechanics do not reflect a real-world phishing detection task (i.e., with full email content under time pressure).

2.1.3 Human-Game Interaction Mechanisms in Anti-Phishing Games

Game mechanics are the actions through which a player interacts with the state of the game [73], and it is through these mechanics that teaching methods are translated into gameplay [69]. Most anti-phishing games "rely on very similar game mechanics" and "do not allow detailed assessments of players' acquired knowledge" [68]. Storylines in serious games have been shown to boost learner engagement [60]. Yet, most anti-phishing games still lack a storyline [69]. Non-digital games, like board games, are often based on roleplay mechanisms [69]. This design choice is also reflected in digital anti-phishing games such as *What.Hack*, where players roleplay as a bank employee classifying emails [90].

Interactive games have been found to be more effective than training learners through traditional learning material [72, 90]. Most games teach concepts separately from game mechanics [69], creating a separation between learning and gameplay. Roepke et al. [68] was an exception, having players create phishing URLs through interactive game mechanics. A focus group found that players required reward mechanisms as incentives and for greater engagement [57]. Furthermore, they wanted to track their progress and know how far along they were in completing the game [57].

Participants in a study requested a competitive multiplayer mode in anti-phishing games [68]. Gamification research has found, however, that competition does not necessarily improve learning outcomes [85], especially compared to collaboration [33]. Digital games need to compensate for the lack of in-person knowledge transfer from an expert in a scalable manner [69]. Anti-Phishing Phil [72] had a non-player character (NPC) who provided hints during gameplay. *What.Hack* let players request help from an NPC who would reveal correct answers [90]. Similarly, Roepke et al. [68] used an NPC tutor to help users learn the game mechanics and develop phishing detection skills.

Though Roepke et al. [68] found that richer game mechanics did not improve learning outcomes compared to earlier games, they suggested that in-game log data from the usage of these mechanics could be used to model player behaviour. Since higher workload leads to higher time pressure which then increases the likelihood of being phished, Wen et al. [90]

suggested that games should reflect this real-world time pressure component when reading emails. Such effects suggest that phishing email detection depends on how a person cognitively processes the email content itself.

2.2 Persuasive Message Processing

The way in which a person cognitively processes a phishing email is a key determinant of whether they will fall for it [86]. Gavett et al. [32] tested five executive function measures against phishing susceptibility, finding that cognitive ability predicts detection performance differently by age. In cognitive science, dual process theories suggest that some problems we encounter trigger a rapid answer while others require deliberation [63]. Rapid answers are essentially autonomous, while deliberate answers require access to our working memory [63, 81].

For example, certain properties of an email, like the sender name matching our friend's name, could trigger a rapid "muscle" memory rather than deeper thinking. Researchers are still unclear how these two forces interplay when we make decisions [63]. Pennycook [63] suggests that cues present in a problem trigger multiple answers in real-time, and when there is a conflict between these answers, deliberation is used before responding.

The Heuristic-Systematic Model (HSM) suggests that persuasive messages can be processed heuristically or systematically [17]. Various factors, such as contextual cues and time pressure, can influence which processing mode is activated [17, 81]. In the heuristic mode, an individual relies more on surface-level cues in the message with less cognitive effort, whereas in the systematic mode, they pay closer attention to the message content itself [17]. In the context of phishing, heuristic processing means skimming the email, while systematic processing involves deliberating on the body of the email and connecting it to prior knowledge, yielding better detection outcomes [86].

Vishwanath et al. [86] extended the HSM in the context of phishing in the form of the Suspicion, Cognition, and Automaticity Model (SCAM). They found that heuristic processing leads to lower suspicion of phishing emails [86]. An individual may process emails

in the heuristic mode because of routine or habitual email use [86]. The ineffectiveness of traditional anti-phishing learning material could be explained by the models of persuasive messaging that occur when a learner encounters a phishing email in the wild. When a user encounters an email under certain circumstances, like task overload, they would peripherally process the email and may depend on surface-level heuristics to make their phishing or safe classification. Therefore, research points toward the fact that anti-phishing training needs to account for these varying modes of message processing.

2.3 Learning Theory in Anti-Phishing Game Design

A theoretical foundation of learning is required to create games that produce learning beyond just entertainment [49, 84]. Participants in a study noted that they wanted feedback inside the game environment itself, rather than as a separate module [57]. Formative feedback benefits learning, especially when it is delivered as part of the learning activity [41]. *What.Hack* [89] provided immediate feedback during email classification for reflection on mistakes. Similarly, *Anti-Phishing Phil* was designed to encourage players to reflect on their mistakes during the game [72].

As discussed in Section 2.2, learning material should simulate the time pressure that drives users toward heuristic behavior and ensure that their learning would be adaptable to these conditions. Sarno and Neider [70] found that email classification times were dependent on the prevalence of phishing among the email samples present and that high task load can reduce classification accuracy. Bergh et al. [11] showed the simulated stress through enforcing a strict 20-second time limit on email classification decreased phishing classification performance. However, no in-depth analyses of the effects of time pressure through designing in-game interaction mechanics to simulate this effect were found in the surveyed literature. Additionally, established metrics that measure this balance between accuracy and time, such as Inverse Efficiency Score (IES) [82, 83], have similarly not been adopted as performance measures of learning in anti-phishing games.

Collaborative learning has been shown to improve test scores and knowledge reten-

tion over individual instruction [46]. Tchakounte et al. [79] recommended a collaborative cloud-based anti-phishing game design, arguing that games with multiple players would improve learning outcomes. Despite these benefits, few cybersecurity games have explored collaborative play [48]. A review of eight anti-phishing games found that only one supported multiple players [79]. Hafner et al. [34] developed a Dungeons & Dragons-inspired multiplayer board game, Tabletop As Social Engineering Prevention (TASEP), to teach cybersecurity practices, evaluating it through user feedback, but without pre/post learning measures or a comparison condition.

Katsantonis et al. [48] propose a collaborative board game called *Yellow Training*, in which players work together to defend their organisation against malicious attackers. *Yellow Training* similarly measured qualitative aspects of the game, like recorded group discussions, without a control condition or pre/post learning assessment. This highlights the need for a study that compares collaborative anti-phishing gameplay against a control condition, with pre/post learning measures of phishing detection performance.

However, these existing games are board games that require physical game material and an expert to be present for each session. In comparison, digital games are not limited by these constraints and can scale remotely to any number of concurrent sessions. A meta-analysis of 45 studies found that online instruction was as effective as in-person modes for improving learning outcomes [56]. The study also found that blended instruction, i.e., hybrid instruction that combines in-person and online, was more effective than both, with the confounding caveat that these studies also contained more learning resources and activities [56]. These findings suggest online or hybrid education as a valid choice for scaling anti-phishing games, but whether such games are as effective as in-person ones remains unclear.

2.4 Influence of Motivation and Behavioral Psychology in Learning

The Fogg Behavior Model [30] suggests that a behaviour is more likely to occur when an individual has the ability to perform it, sufficient motivation to do so, and a trigger to initiate it. Therefore, anti-phishing training should focus not only on the learning itself, but must provide the skills needed to deploy the material learned when the need arises during regular email workflow. Specifically, the Fogg Model suggests this requires increasing motivation and self-efficacy for the task at hand, and providing triggers that push the user towards more deliberation when a phishing email is encountered.

Chen et al. [19] found that both group discussion and role-play based training improved perceived self-efficacy, i.e., an individual's belief in their ability to perform a task [10], and increased phishing reporting after training ($n = 105$). The Extended Parallel Process Model (EPPM) proposes that when perceived threat is high but a person's perceived efficacy is low, people respond based on fear [91].

Phishing training should therefore build self-efficacy alongside threat awareness, or learners will continue to get phished. Shahbaznezhad et al. [71] described self-efficacy in the phishing context as the user's perceived ability to avoid falling for a phishing attack. However, they found that self-efficacy did not significantly predict phishing avoidance, despite its importance as stated in the literature, suggesting that other traits may moderate its effect.

Motivation can be sub-divided into two types: intrinsic and extrinsic motivation. While intrinsic motivation is affected by internal factors like desire for self-improvement and curiosity, extrinsic motivation is driven by external factors like grades, rewards, and competition [64]. The MSLQ is a survey of 81 items that measures motivation through scales such as intrinsic and extrinsic goal orientation, task value, self-efficacy, and test anxiety, as well as learning strategies such as critical thinking, effort regulation, and metacognitive self-regulation [64]. Participants answer each item on a Likert scale ranging from 1 (not at all true of me) to 7 (very true of me).

Researchers from different domains (e.g., education and healthcare) have used the MSLQ survey to measure motivation [39]. The MSLQ is flexible and modular by design, and can be adapted for the particular requirements of a specific study [64]. However, the exploration of the MSLQ in anti-phishing training has been restricted to a small subset of 8 adapted questions in [76], possibly constrained by the length of the questionnaire. To address this, we first study the MSLQ in an undergraduate engineering design course across different instruction modes (in-person, hybrid, and remote) in Chapter 3. After establishing the validity of the MSLQ in a university-oriented organisational setting, we derive a reduced latent construct, namely MPA, which we subsequently apply to our anti-phishing training domain.

In the Theory of Planned Behavior, subjective norms describe whether individuals feel socially pressured to engage in or avoid a behaviour [3]. In the context of phishing, Shahbaznezhad et al. [71] captured subjective norms as the expectations of organisational entities like supervisors, colleagues, and the IT department. Higher subjective norms significantly reduced the likelihood of clicking on phishing emails [71]. However, this social pressure can be exploited: Alyahya and Weir [7] found that participants with higher subjective norm values were more likely to engage with phishing emails with social components. A large study of working adults ($n = 985$) found that the influence of social proof, i.e. following the actions of others, was a significant predictor of falling for phishing emails [62]. These findings motivate our analysis of subjective norms as a predictor of phishing detection learning after training with *Phooled!*.

2.5 Predictors of Phishing Susceptibility

Individual characteristics such as personality traits, cognitive impulsivity and security awareness are factors in phishing susceptibility [15]. Halevi et al. [35] found that employees who believed they were safe from phishing were in fact more likely to get phished, suggesting that awareness about phishing attacks does not reduce the chances of a person getting phished. Moreover, compliance intention was not found to be correlated to compliance behaviour for phishing avoidance [42].

Crucially, Downs et al. [26] found that perceiving dire consequences for getting phished does not improve detection. In fact, Canfield et al. [16] found that it leads to misclassifying safe emails as phishing, increasing false positive rates. This was confirmed by Zielinska et al. [92], where fear-based anti-phishing training helped as well as non-fear-based education, but at the cost of a rise in false positives after training.

A user's traits such as curiosity, risk-taking, and Internet anxiety increased phishing susceptibility [58]. Jalali et al. [42] found that the workload of an employee increased the likelihood of getting phished. Abroshan et al. [2] found that anxiety and stress increased the odds of falling for a phishing attack. Participants' stress levels were also found to reduce their phishing detection accuracy [11]. Consistent with these findings, mindfulness training reduced the chances of getting phished in [45]. Therefore, it is important to study factors such as self-efficacy and test anxiety, both of which are scales in the MSLQ survey [64]. In Chapter 5, we corroborate these prior findings, with motivation and perceived ability predicting better phishing detection performance and test anxiety predicting worse performance under specific conditions.

2.6 Determinants of Phishing Email Detection Difficulty

Prior research suggests that, apart from individual characteristics, the features of an incoming email itself also determine the difficulty of detecting phishing. For example, emails originating from a known sender, i.e. a known *from* address or *from* name, increased the likelihood of getting phished [58]. Phishing cues or attributes present in the email message are used by the user to examine incoming emails [16].

Phishing cue detection is based on previous exposure to these cues or knowledge about them [16]. These cues include suspicious URLs, requests for urgent action, and grammatical errors [27]. Musuva et al. suggest that peripheral cues, such as source credibility and appearance, influence the way that we interpret a message [59]. These cues reduced participants' ability to detect phishing threats [59]. Interestingly, the quality of the argument

used by the attacker in the message (e.g., how persuasive the phishing email is) did not make phishing detection more difficult [59].

Vishwanath et al. found that users fall for phishing because they peripherally process cues in the email [87]. Additionally, they found that urgency cues diverted attention away from other phishing cues present in an email [87]. Despite grammar and spelling being claimed as phishing cues, Butavicius et al. showed no link between these cues and increased detection [14]. Their study also showed that the more deliberate action of checking the URL of a link before clicking it was linked with better phishing detection [14]. However, 34% of participants did not examine URLs, and even when they did, they only examined 14% of them [14], suggesting that they primarily opted for a more heuristic approach.

The role of user demographics in phishing susceptibility has been widely studied, though findings vary considerably across studies. For example, Kumaraguru et al. found that younger participants were more susceptible [50], while Baillon et al. found older employees were more susceptible in a large field experiment [9], and others found no significant age effect [32,42,58]. Similarly, gender was not found to be significant in most studies [32,52,58], though Marriott found that women performed worse [53]. Education level showed a similar pattern, with Gavett et al. finding that higher education reduced susceptibility [32], Jalali et al. finding no significant effect [42], and Marriott finding that higher education was associated with worse performance [53]. These contradictory findings suggest that user demographics are poor predictors of phishing susceptibility.

2.7 Research Questions

As described in the preceding sections, our literature review identified several gaps in the anti-phishing training domain. Specifically, most existing anti-phishing games focus on URLs, not the broader features of the email itself [69]. Moreover, to our knowledge, no digital anti-phishing game has compared Single-Player and Collaborative training [48, 79]. While individual characteristics such as motivation and subjective norms have been studied as predictors of phishing susceptibility, whether they predict phishing detection

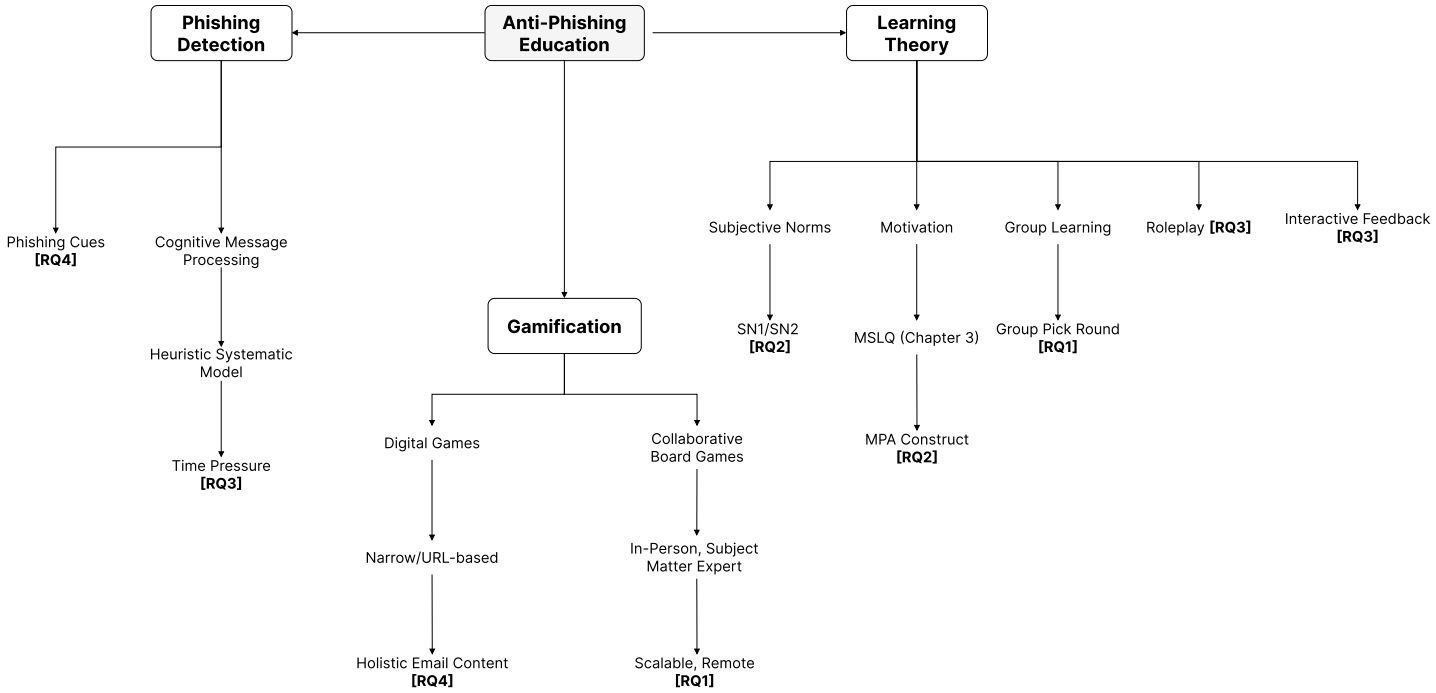


Figure 2.1: Theoretical foundations of our approach and their mapping onto our research questions

improvement through game-based training has also not been examined.

Additionally, whether the MPA scale, a reduced construct derived from the MSLQ in Chapter 3, is sufficient to predict phishing detection improvement in an anti-phishing training context remains an open question. Furthermore, whether the instruction mode of game-based training, such as in-person or remote delivery, impacts learning or influences these individual characteristics has not been examined either. To explore these gaps, we propose four Research Questions (RQs) below. RQ1 and RQ2 are confirmatory in nature and will be tested using specific hypotheses as described in Section 4.2, while RQ3 and RQ4 are exploratory.

RQ1 Are collaborative anti-phishing training games more effective than single-player games at improving users' phishing detection?

RQ2 Do individual characteristics, such as motivation and subjective norms, predict users' improvement in phishing detection after training with anti-phishing games?

RQ3 How do anti-phishing training in-game interaction mechanics, such as time pressure, interactive feedback, and roleplaying, affect phishing detection performance and improvement?

RQ4 How do email features, such as email sample type and phishing cues, influence the difficulty of phishing detection?

Chapter 3

Exploration of Motivation and Learning Strategies for Education

In this chapter, we present a study of motivation and learning strategies in a first-year engineering design course at the University of Ottawa [24]. The data was collected in collaboration with the course instructors. In this study, we use Pintrich's Motivated Strategies for Learning Questionnaire (MSLQ) [64] to measure motivation across 155 students in five cohorts taught under varying instruction modes, namely, in-person, Hybrid Flexible (HyFlex), and online. We identify a latent variable named "Motivation and Perceived Ability" (MPA), which reduces the extended MSLQ questionnaire into a construct of its subset. This MPA scale is later used to measure motivation in Chapter 5.

3.1 Introduction: Intrinsic and Extrinsic Motivation

Intrinsic motivation can be viewed as the set of internal forces that drive people to behave in different ways and is the result of both the personality and the abilities of the individual, combined with their previous experiences [25]. These intrinsic motivation factors co-exist with other (extrinsic) motivating factors, such as grades, for students. Measuring moti-

vation is difficult, because it varies with the individual and can even vary with a specific individual over time [37, 54].

Further, subtle extrinsic factors that are present in the environment influence both types of motivation. Indeed, the effects of such factors can be hard to separate from each other, too, when self-assessment tools are being used. Despite these concerns, we use self-reported data to identify the significance of intrinsic motivation, as it affects activities in a team-based, project-based engineering design course for first-year students. We want to determine whether other factors associated with near-"authentic" (i.e. a design problem with a real-world client) engineering design tasks are correlated with motivation scales such as intrinsic motivation. For this purpose, we use a survey mechanism, initially created by Pintrich et al., called the "Motivated Strategies for Learning Questionnaire (MSLQ)" [64].

MSLQ attempts to quantify the levels of both intrinsic and extrinsic motivation and learning strategies in subjects and was originally defined specifically for an academic environment. MSLQ relates motivation to higher cognitive functions (e.g., those relating to organisation, peer learning, critical thinking, as well as a few others). The survey tool has been adapted for other environments [18, 22] but has been criticised by researchers and shown to be inaccurate or unreliable in certain situations [18, 39, 67]. Nevertheless, the method continues to be used, and this persistence is in part because the results can be analysed at different levels, as will be explained in more detail later.

Ethics approval for this research project was duly obtained from the University of Ottawa Research Ethics Board (REB), under file number H-02-24-10020.

3.2 Description of Experimental Method

During this study, in the GNG1103 - "Introduction to Engineering Design" course at the University of Ottawa, students were taught engineering design in three concurrent course sections, each led by a different instructor. Data was gathered over five distinct terms with different cohorts of students. The course is offered in both English and French.

In the Fall 2021 term, there were three different instructors: A, B and C, with class sizes that varied from 63 to 84 students, depending on the specific section. For the Winter and Fall 2022 terms, instructor C was replaced by a new one, D, to create a new set of instructors: A, B and D. Replacing the third instructor each time, the 2023 Winter term instructor set became: (A, B, E) and then changed to (A, B, F) in the final 2024 Winter term. The details of the course sections are given in Table 3.1.

Table 3.1: Nature of Learning Environment

Term	**Instructors**	**Instruction Mode**		
		Sync. Lectures	**Labs**	**Group Project**
2021 Fall	ABC	Online	Mostly Online	Online
2022 Winter	ABD	HyFlex	HyFlex	HyFlex
2022 Fall	ABD	Mostly In-Person	Fully In-Person	Fully In-Person
2023 Winter	ABE	Fully In-Person	Fully In-Person	Fully In-Person
2024 Winter	ABF	Fully In-Person	Fully In-Person	Fully In-Person

All course assessments, lecture content, and lab content are kept identical for students in each of the different sections, although there are natural variations in style and delivery with different instructors. There is a single final exam for all course sections. The design project topic is a continuous project throughout the term in which the students work with a client to address real-world problems. For example, the project could be designing an automated farm weeder or a robot to provide guided museum tours.

The design project topic varies with each course section, but is identical for all students in a particular course section. Within a specific course section, project teams are organised into groups of four to five students, to work on the project together. This structure allows students to collaborate with their peers to enhance their learning experience. A student project manager and a graduate student teaching assistant are assigned to each team. The project manager, a senior student, who has already taken the course, actively engages with the team to offer design feedback and to provide mentorship for the management of the

team's schedules and timelines. The teaching assistant offers general evaluative feedback regarding the team's overall progress and assesses their deliverables, such as their weekly group assignments.

During the pandemic, all students in each section wrote a final exam at the same time, with questions drawn randomly from large question pools, resulting in no two exams being exactly the same. This randomisation was carried out in an attempt to reduce plagiarism. A significant portion of a student's grade for the course (45%) is derived from their performance on the group-based project deliverables, while the rest is determined from individual assessments (e.g., exams, quizzes, reflective essays, etc.). There are a total of five quizzes in the course, distributed throughout the term. A portion of each quiz is a formative take-home essay assessment, where students are allowed to access external resources in a lower-stress environment. The final exam is structured in the same way as these quizzes and functions as a summative assessment performed "live" using a Learning Management System (LMS).

Five distinct cohorts of students were surveyed for this research study. The first cohort of three course sections of students attended lectures synchronously, working in groups in an environment that was completely online (because of the COVID-19 global pandemic) in the Fall of 2021. Students were surveyed after the course was complete, but before final grades were announced.

In the Winter of 2022, except for the very first month of the term, which was also held completely online, the second cohort of three course sections of students studied and worked in a HyFlex learning environment. In the HyFlex offering of the course, students could choose to attend either "in person" or online for any or all the course components (i.e. lectures, labs, or group project work). The instructors taught a concurrent mixture of online and in-person students in their lectures. The Teaching Assistants provided the same kind of supervision and instruction during the lab sessions.

In both online and HyFlex terms, students were asked to attend lectures synchronously, but had the option of watching a video recording of the lecture later, asynchronously. Although asynchronous attendance was possible, it was actively discouraged, given the

large amount of interactive group activity built into the course lectures, labs, and group work. Online versions of all labs were provided for the entire durations of both online and HyFlex terms. In addition, course projects were selected with a greater software component than was typical, in terms prior to 2020.

The survey was repeated in the 2022 Fall term, where the mode of instruction was "mostly in-person". While the primary mode of instruction was synchronous and "in-person", a few students, forced to remain abroad due to the pandemic, were given the opportunity to attend lectures online. The remaining course components, other than the lecture, were "fully in-person". The final two terms, the 2023 Winter and 2024 Winter terms, had all of their course components "fully in-person".

A "continuous improvement" process has been used to optimise GNG1103 course content, since its inception in 2016. However, by 2020 90% of the lecture content and course structure was stable with minor modifications being made, on a term by term basis. Interestingly, the COVID pandemic created some interesting new possibilities, in terms of computer-based collaborative work, but many or most of these were replaced, in turn, with more "live" alternatives, as pandemic restrictions were eased.

All questions were identical in the surveys for all cohorts and across all course sections. We measure the levels of both student motivation and learning strategies under the different online/HyFlex/in-person working conditions for each set of students. We analyse the data across sections, looking for statistically significant differences in the results, based on the specific section instructor and/or project topic (Table 3.1). Initially, data was collected as part of the "continuous improvement" effort, but ethical review and approval was obtained in the final year for the "secondary purposes" outlined in this chapter.

3.3 Survey Results

A standard MSLQ survey consists of 81 survey items (i.e., questions) and is accompanied by a "manual" [64]. The items are assessed using a 7-point Likert scale that spans from "not at all true of me" to "very true of me". The questions are grouped into a set of

15 “Scales”. For example, one of these scales is ‘intrinsic motivation’. A smaller set of “components” is then also defined, as a super-set of logically related groups of those same scales. For example, “extrinsic motivation”, “intrinsic motivation” and “task value” scales are all then grouped under a higher level “Value” component categorisation. This categorisation allows for a higher level of analysis to be performed.

The MSLQ survey is notably comprehensive due to its extensive set of 81 questions. The estimated time to complete the survey is 15–20 minutes. A total of 167 students started our survey; however, only 155 individuals successfully completed it, providing valid responses to all of the questions. The 7.19% dropout rate includes those students who provided incomplete responses to any of the 81 items on the Likert scale. All such data were excluded from our analysis. The participation rate in the student survey varied between 18% and 26% of the total class size.

The Shapiro-Wilk test was performed for each scale, and the results indicated that the data followed a normal distribution. The median values of the entire cohort, taken over all terms, are presented in Figure 3.1a. Overall, students had above-average scores on the motivation and learning strategy scales. The notable exception was the peer learning scale, with a mean of 3.363 (SD = 1.363). Interestingly, this result is consistent with Pintrich’s original work [64], but is of concern, given the intended peer-based, project-based, learning model for the first-year engineering design course.

To observe differences between the three modes of instruction, we perform a one-way Analysis of Variance (ANOVA) test, with statistical significance evaluated at $\alpha = 0.05$. This allows us to observe any variations in the MSLQ scales across the various terms. Table 3.2 reports the sum of squares (SS), F-statistic, and p-value from the one-way ANOVA for each MSLQ scale across the five terms. The ANOVA test, performed across terms, for the peer learning dependent variable was not significant, which means that remote learning does not explain the discrepancy in the mean value of peer learning. Previous studies [88] also observed low values for peer learning, compared to other MSLQ scales. Therefore, this pattern might be a defect/feature of the tool or a larger trend in education, rather than being related to specific variation observed in our study (e.g., because of COVID) with first-year engineering design students.

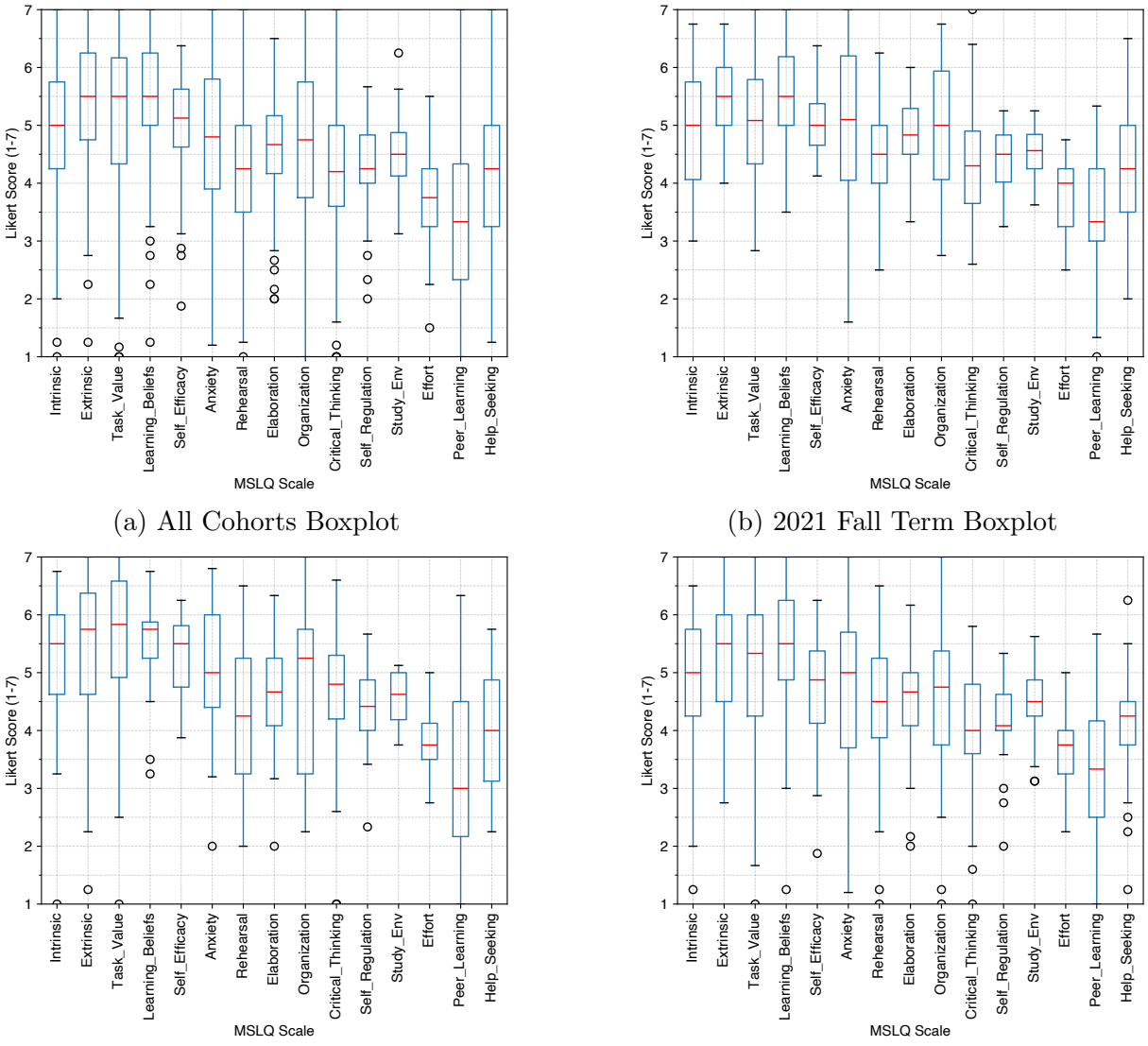
Table 3.2: ANOVA Results for Various MSLQ Scales and Components

Component	**Scale**	SS	F	p
Value	Intrinsic Motivation	5.733	1.101	.358
	Extrinsic Motivation	2.898	0.620	.649
	Task Value	10.847	1.584	.181
Expectancy	Learning Beliefs	2.238	0.526	.717
	Self Efficacy	7.938	3.291	.013*
Affective	Test Anxiety	14.348	2.007	.096
Cognitive & Metacognitive Strategies	Rehearsal	7.110	1.177	.323
	Elaboration	4.604	1.626	.171
	Organisation	5.875	0.891	.471
	Critical Thinking	6.948	1.258	.289
	Self-Regulation	3.331	2.189	.073
Resource Management Strategies	Study Environment	0.491	0.463	.763
	Effort Regulation	0.983	0.556	.695
	Peer Learning	0.407	0.055	.994
	Help Seeking	2.531	0.486	.746

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

3.3.1 Correlation of MSLQ Scales

The correlation between the different MSLQ scales is assessed to analyse the relationship between the various aspects of student motivation and learning strategies within the context of the GNG1103 design course. Cronbach's alpha (α) measures internal consistency and was calculated for the entire dataset, for all cohorts. The resulting value of 0.854 indicated good internal consistency [78]. We calculated the pair-wise Pearson correlation coefficient (r) between each MSLQ scale, with statistical significance evaluated at $\alpha = 0.05$. Figure 3.3



(a) All Cohorts Boxplot

(b) 2021 Fall Term Boxplot

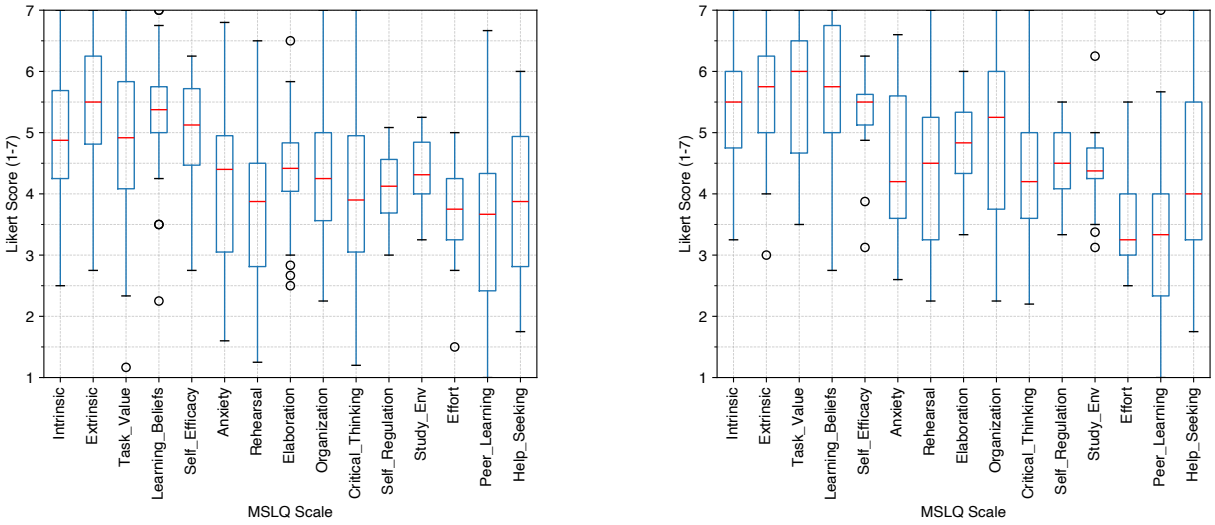
(c) 2022 Winter Term Boxplot

(d) 2022 Fall Term Boxplot

Figure 3.1: Term-wise Boxplots of MSLQ Scales for All Cohorts, 2021 Fall, 2022 Winter, and 2022 Fall terms (n = 155)

shows the correlation matrix, with coefficients labelled only when both $p < 0.05$ and $r > 0.5$ (a strong effect size), but colour is used to indicate the specific value regardless.

We categorised correlations according to established guidelines defined by [23]. The



(a) 2023 Winter Term Boxplot

(b) 2024 Winter Term Boxplot

Figure 3.2: Term-wise Boxplots of MSLQ Scales for 2023 Winter and 2024 Winter terms (n = 155)

strongest correlation was between Intrinsic Motivation and Task Value, representing a strong correlation. In contrast, Extrinsic Motivation, Test Anxiety, Critical Thinking, Study Environment, and Effort Regulation showed no significant correlations with any other scale.

The Intrinsic Motivation scale was found to be correlated with the measure of Task Value ($r = 0.79$), which might not be surprising. This strong correlation indicates that students who have a high intrinsic goal orientation found the engineering design course material to be of high interest and utility, participating in course activities out of curiosity and/or a desire for mastery [64], rather than due to external factors, such as grades. Originally, in Pintrich's work, these scales were compared with graded outcomes only, rather than as a correlation between the scales.

Furthermore, these students had a moderate positive correlation with their Control of Learning Beliefs and Metacognitive Self-Regulation scales. This moderate correlation implies that such students felt that their outcomes in the course depended on factors that they could control. Such controllable factors might be: studying habits, planning,

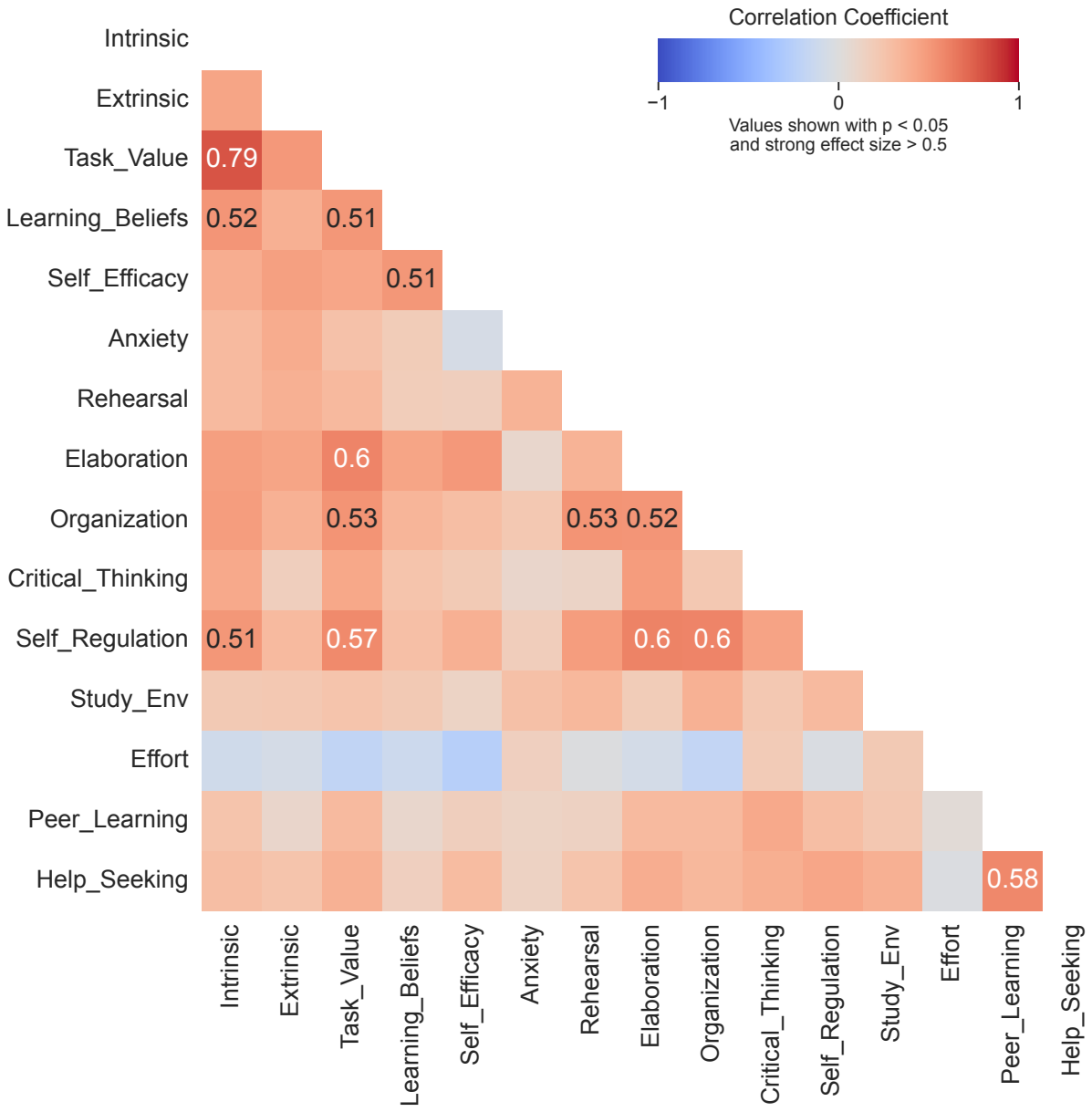


Figure 3.3: Overall Correlation Matrix for MSLQ Scales

monitoring, and regulation, as well as other existing skills, related to self-control.

Students who scored highly on the Metacognitive Self-Regulation scale were also more

likely to be organised in their study methods and employ elaboration strategies, such as summarising readings in their own words or relating course material to other courses they were taking [64], during their study. They believed that they had built systematic connections between the knowledge that was being absorbed, rather than purely reciting back course material. Since the course content was identical across all sections, this correlation should not vary by semester or by instructor, but we will analyse this pattern in the following sections. For each term in the study, the medians, inter-quartiles, minima, maxima, and identified outliers are shown in Figures 3.1 and 3.2.

In this study, all of the first-year engineering design students, who valued the course highly, were more likely to use elaboration strategies for studying, rather than just using simple rehearsal methods. Unsurprisingly, students who prefer to study in groups with their peers were also more likely to seek out help. This help could come either from those other students or from the instructor, or from a senior student (e.g., a teaching assistant or student project manager) when they faced difficulty with the learning material or with the implementation of it as a final design prototype.

3.3.2 Differences Between Modes of Instruction

The value of Cronbach's alpha (α) was calculated individually for each term and shown in Table 3.3, where values above 0.70 are generally considered acceptable [78] and values above 0.60 are considered acceptable for exploratory research [47]. Although the Fall 2021 term had a slightly lower α value of 0.674, this exceeds the exploratory threshold. All subsequent terms achieved strong internal consistency above 0.8.

Indeed, examining Figures 3.1 and 3.2, the average values are remarkably consistent across all terms (i.e., all three instruction modes and all six instructors). A one-way ANOVA test was performed across all five terms to observe significant variations in means for all scales to observe differences between the modes of instruction. The ANOVA test was performed with an alpha value (α) of 0.05. The results from this test are presented in Table 3.2. A couple of results are more significant, based on their F-Score and corresponding p-value.

Table 3.3: Cronbach's Alpha (α) for each term

Term	**Cronbach's Alpha (α)**
Fall 2021	0.674
Winter 2022	0.871
Fall 2022	0.834
Winter 2023	0.862
Winter 2024	0.904

Most notably, self-efficacy varies significantly across terms, with an F value of 3.290 and a p-value of 0.012. Tukey's HSD test was used to find the semesters of significant variation for the self-efficacy scale. The results are plotted in Figure 3.4. The confidence interval was set to 95%.

Figure 3.4 shows the mean self-efficacy score for each term with 95% confidence intervals from the Tukey HSD test. The only pairwise comparison to reach significance was between the 2022 Fall and 2024 Winter terms, with a mean difference of 0.634 ($p = .013$, 95% Confidence Interval (CI) [0.094, 1.173]). This likely reflects students' increased self-efficacy once in-person instruction had normalised after the disruptive transition from remote to in-person during the Fall 2022 term.

Self-efficacy rose during the transition from the online to the HyFlex term, but fell sharply in the transition to the in-person term, before rising again in subsequent terms. Students seem to be relatively confident in their ability in the online term. There is an increase in self-efficacy in the HyFlex term, perhaps due to increased familiarity with the remote learning environment.

However, there is a drop in self-efficacy in the Fall 2022 term, during the transition back to fully in-person instruction modes. As modes of student instruction switched back to being fully in person, students seemed to have been able to regulate themselves better, and the change in the correlation value between self-efficacy and self-regulation is significant. This trend is most noticeable between the cohort of students who were just starting at

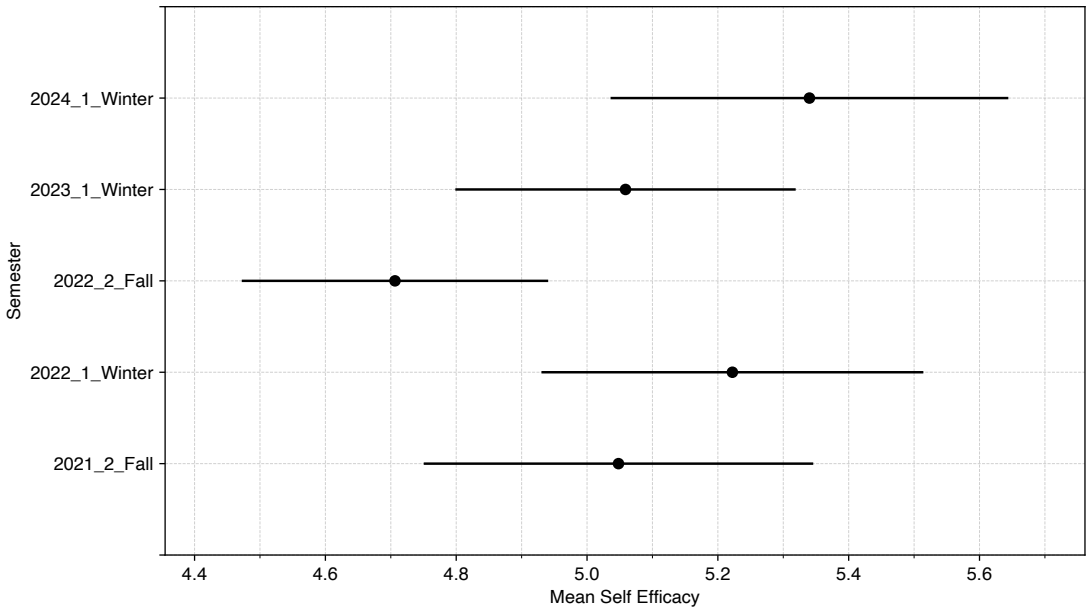


Figure 3.4: Tukey's HSD Test for Self-Efficacy, by term, with confidence interval for each mean

university, having done their final years of high school with online learning environments, and those who had already completed one university term in person. Perhaps students were then confident in their abilities in the now-familiar in-person mode of instruction, as opposed to being in a transition state of "mostly" in-person instruction, as the world emerged from the COVID-19 pandemic. The group-based course organisation (as well as the focus on "hands-on" learning in the project and labs and lectures) of the first-year engineering design course could have accentuated these differences even more.

The improvement in self-efficacy between the 2021 Fall and 2022 Winter terms is interesting. In fact, the levels in the 2022 Winter term are higher than those in the two subsequent terms, each of which had in-person formats. If subsequent terms had also been conducted online, the trend of self-efficacy might have continued to increase, but there is no way to confirm this trend. The one-way ANOVA test for the self-regulation scale was found to have a p-value of 0.073, close to the specified significance level of 0.05. The Tukey HSD test was performed for this scale, but no significant differences were observed between

these other terms.

To examine the differences in correlation across the different modes of instruction, the correlation matrix was calculated for each individual term. The correlation matrices were transformed using Fischer's Z transformation, to normalise the distribution. For each pair of terms, the corresponding MSLQ scales in the Z-score matrices were compared, as shown in Figure 3.5. A critical value of ± 1.96 was considered for analysis, which corresponds to a two-tailed p-value of 0.05. A positive Z-score for a pair of terms indicates that the correlation in the first term was greater than the second. Larger Z-scores indicate greater difference in their correlations.

Fischer's Z-test revealed statistically significant differences in the correlation between the Intrinsic and Extrinsic motivation scales across terms. Specifically, the Z-test values were -4.994 (2021 Fall and 2022 Winter), 3.77 (2022 Winter and 2022 Fall) and -2.019 (2023 Winter and 2024 Winter terms). The correlation spikes at 0.85 for the 2022 Winter term. The correlation declines sharply to 0.28 during the transition to the in-person mode of instruction during 2022 Fall. The correlation gradually increased across subsequent in-person terms, reaching 0.38 in 2023 Winter and reaching a high value once again in 2024 Winter. This suggests that intrinsic and extrinsic motivation reinforced each other during the HyFlex term, whereas general uncertainties during the transition periods induced a decoupling of these scales, perhaps due to ambiguity of assessments and change in environment.

A comparable, but less pronounced, trend was observed between the Task Value and Extrinsic Motivation correlation pair. This pair describes the students who found the course to be valuable, along with also needing external factors, as a source of motivation. For this pair, the significant Z-test values were -3.382 (2021 Fall and 2022 Winter) and 2.048 (2022 Winter and 2022 Fall). These values reflect a sharp increase in correlation during the HyFlex term from 0.06 to 0.78, and a subsequent decline to a moderate value of 0.48 in the in-person term. The value remained relatively constant during the in-person terms.

The Control of Learning Beliefs and Extrinsic Motivation correlation pair reflects stu-

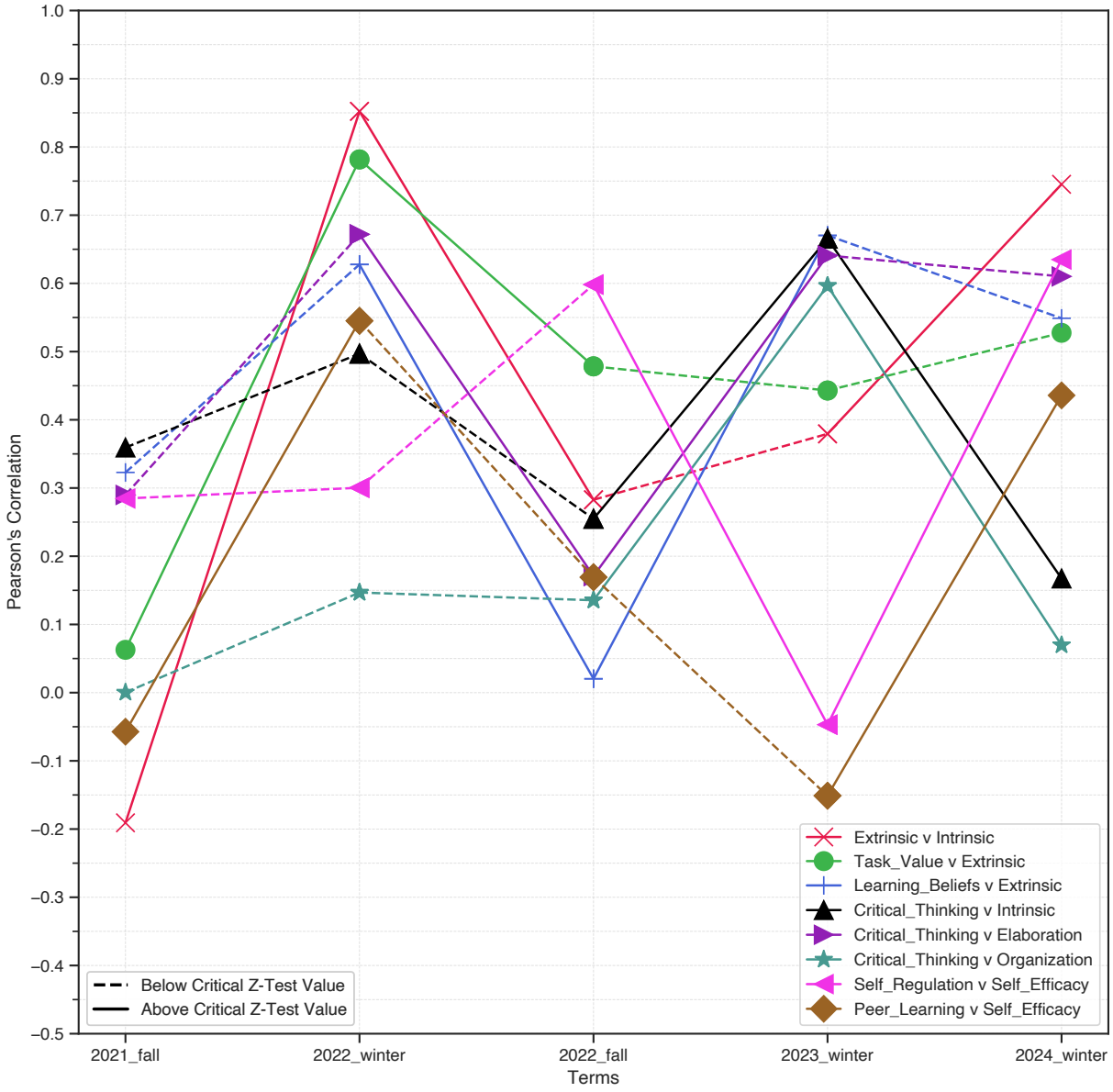


Figure 3.5: Significant Fischer's Z-Test Values for MSLQ Scale Correlation Pairs Plotted Term-Wise. Solid lines denote periods with critical Z-Test Values (see legend).

dents who are motivated by external factors, but who also believe that their efforts in the course will be meaningful. The significant Z-test values were 2.778 (2022 Winter and 2022

Fall) and −3.306 (2022 Fall and 2023 Winter). This observation is caused by the fact that the high correlation value present in 2022 Winter decreased significantly during the 2022 Fall transition term, before significantly increasing again in the subsequent term. In the 2022 Fall transition term, the correlation between Extrinsic Motivation and Control of Learning Beliefs declined significantly, suggesting that the return to in-person may have introduced uncertainty about whether effort would result in better learning outcomes.

For the Critical Thinking and Intrinsic Motivation pair, the significant Z-test values were −2.266 (2022 Fall and 2023 Winter) and 2.274 (2023 Winter and 2024 Winter). Generally, these two scales are positively correlated with each other, but there were two terms of low correlation, namely, 2022 Fall and 2024 Winter. In these terms, very few intrinsically motivated students also viewed themselves as critical thinkers. The Critical Thinking and Elaboration pair showed significant Z-test values of 2.484 (2022 Winter and 2022 Fall) and −2.453 (2022 Fall and 2023 Winter). This result can be attributed to an outlier of low correlation during the 2022 Fall term (0.171 correlation coefficient), which was the only term without a strong positive correlation. As with self-efficacy, it was observed that the 2022 Fall term bucked the trend and was the only term in which the Critical Thinking scale did not correlate with any other scale among the students. Figure 3.3 shows that intrinsically motivated learners are more likely to use Elaboration as a study method than rehearsal. It is within this cohort of students that the decline in Critical Thinking is observed during the transition term.

The 2023 Winter term saw a significant increase in the Critical Thinking and Organisation correlation pair, with Z-test values of −2.304 (2022 Fall and 2023 Winter) and 2.216 (2023 Winter and 2024 Winter). The 2023 Winter term was the only term with a strong positive correlation for this correlation pair. In contrast, the correlation coefficient for Metacognitive Self-Regulation and Self-Efficacy was −0.05, which was the lowest among all terms. This observation is reflected in the Z-test values of 3.082 (2022 Fall and 2023 Winter) and −2.858 (2023 Winter and 2024 Winter).

The Peer Learning and Self-Efficacy correlation pair had significant Z-test values of −2.292 (2021 Fall and 2022 Winter) and −2.222 (2023 Winter and 2024 Winter). This trend is explained by the moderate positive correlation of 0.55 and 0.44 during the 2022

Winter and 2024 Winter terms, respectively. A similar increase is visible in Peer Learning correlations with both Rehearsal (-2.251) and Elaboration (-2.547) during the shift from online to a HyFlex mode of instruction.

These Fischer's Z-test results suggest that the relationships between motivational scales are not stable across instruction modes. In particular, correlations were generally higher in the HyFlex term and lower in the Fall 2022 transition term. While the one-way ANOVA test revealed consistent mean scale values across terms, the Fischer's Z-test revealed that the relationships between scales changed significantly with instruction mode.

3.3.3 Instructor-wise/ Project-wise Analysis

There were 6 distinct instructors (each leading classes with distinct projects) for this engineering design course, labelled A to F, as shown previously in Table 3.1. To analyse the differences among these instructors, a one-way ANOVA test is performed across all instructors to observe significant variations in the means for all scales. The ANOVA test is performed with an alpha value (α) of 0.05.

The results show significant differences among only one pair of instructors, namely, A and B. The scales of variation are Extrinsic Motivation, Self-Efficacy, Elaboration, and Self-Regulation. In each of these scales of variation, instructor A scored on average 0.56 points above instructor B. There are no significant variations among any other pairs of instructors. Overall, the results suggest that student motivation and learning strategies were largely consistent across instructors, with Instructors A and B being the only exception, though whether this is due to the instructor or the project could not be confirmed with the available data.

3.3.4 Component-wise Analysis

Using the five MSLQ components defined previously in (Table 3.2), namely: Value, Expectancy, Affective, Cognitive, and Resource Management skills, the overall results are

shown in Figure 3.7a. As shown in Figure 3.6, students score higher on what Pintrich [64] calls "Motivation" scales (e.g., the Value, Expectancy, and Affective MSLQ components) than with "Learning Strategy" scales that are part of the Cognitive and Resource Management MSLQ components. Using a Shapiro-Wilk test to check for normality, the Motivation criteria are found to be not normal, with $p < 0.05$. Therefore, the Wilcoxon signed-rank test must then be used to assess whether the values come from the same (non-normal) distribution.

The "Motivation" and "Learning Strategy" scales are significantly different, based on the Wilcoxon test value ($p < 0.05$). This finding is not unique to first-year engineering design students and is similar to the original study by Pintrich, which had students from several disciplines/faculties over all four years at a community college. This suggests that while students in this course are generally highly motivated, their use of active learning strategies such as elaboration and self-regulation is relatively lower.

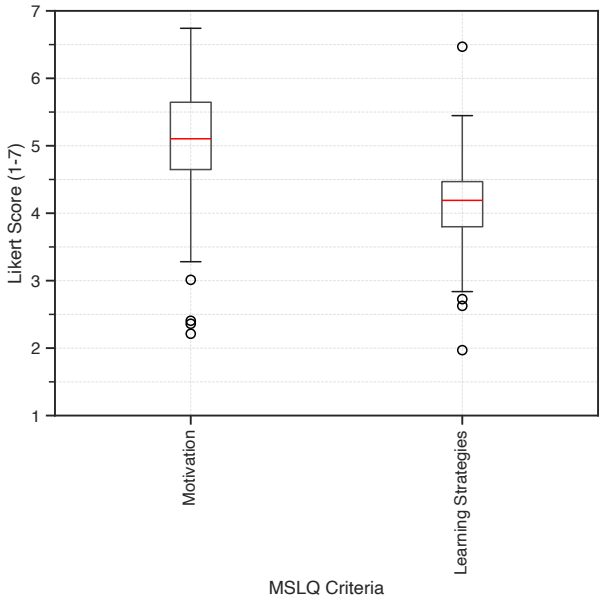
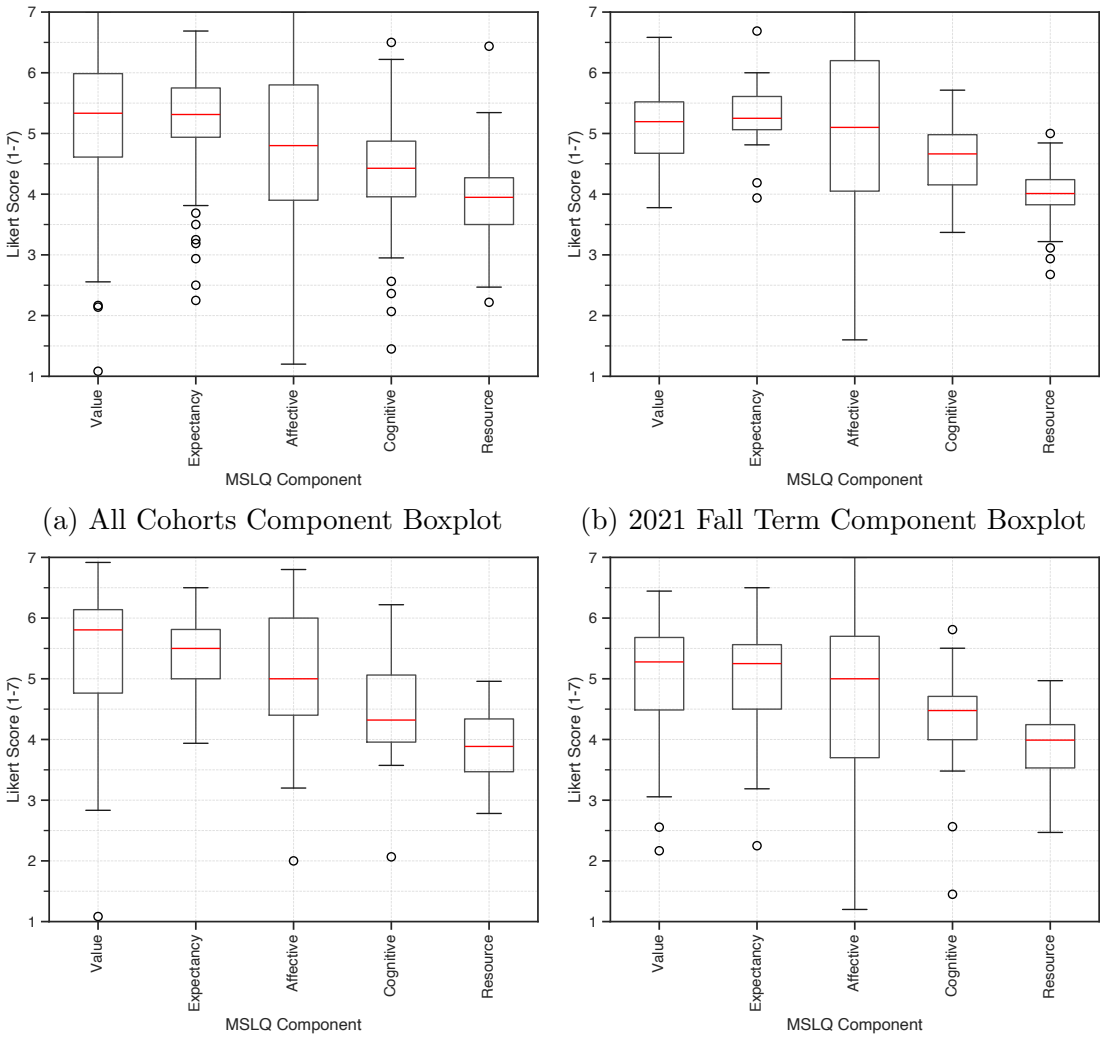


Figure 3.6: Overall Box plot of Motivation and Learning Strategies Criteria

To observe differences between the MSLQ components across terms, a one-way ANOVA test was performed for each component across the various terms. No significant variations



(a) All Cohorts Component Boxplot

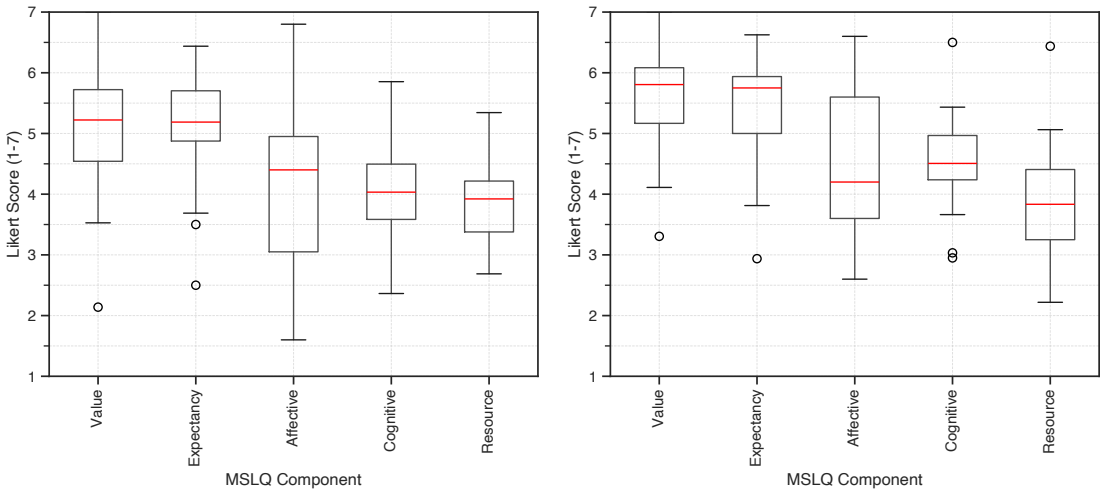
(b) 2021 Fall Term Component Boxplot

(c) 2022 Winter Term Component Boxplot

(d) 2022 Fall Term Component Boxplot

Figure 3.7: Term-wise Boxplots of MSLQ Components for All Cohorts, 2021 Fall, 2022 Winter, and 2022 Fall (n = 155)

were found between the terms, using a p-value threshold of 0.05. However, the Affective component (basically identical to the "Test Anxiety" scale) variance is noticeably larger than the other components, when examined across all terms. The median is consistently near a value of 5, dipping slightly in the two later in-person terms. This observation did



(a) 2023 Winter Term Component Boxplot (b) 2024 Winter Term Component Boxplot

Figure 3.8: Term-wise Boxplots of MSLQ Components for 2023 Winter and 2024 Winter (n = 155)

not seem to be as true in Pintrich's original study, where the range of values is one point or more lower. Almost half of a student's grade in GNG1103 is determined by group-based project-deliverable assessments, but this grading scheme apparently increases "test anxiety".

3.3.5 Exploratory Factor Analysis

We conduct exploratory factor analysis to examine the latent variables present in our dataset. The latent variables in our data are composed of combinations of MSLQ scales. We observe the significance of the factor loadings of these scales and their grouping to form the latent variable. Subsequently, these groupings can be compared with the predefined categorisation of scales as components by Pintrich. A key aim of this section is to determine the significance of MSLQ scales among students to refine learning and course improvement, along with the need to reduce the time-consuming nature of the MSLQ questionnaire, since it will be used to measure motivation during *Phooled!* game play later.

The Kaiser-Meyer-Olkin (KMO) test for sampling adequacy is performed on the MSLQ scale-wise data and is found to be adequate at a value of 0.817. The Bartlett test of sphericity is performed and the data are found to be significant (p-value < 0.05). The scales are passed as observed variables for a Principal Axis Factor Analysis in order to determine latent variables present in the data. An oblique rotation, namely *promax*, is utilised. An oblique rotation is chosen for the initial analysis because the factors are presumed to be correlated.

A total of four factors are found above an eigenvalue of 1.00 in our analysis, and a scree plot is created as shown in Figure 3.9. In Table 3.4, the factor loadings are presented with a threshold of 0.5, in line with the recommendations of [29]. Furthermore, the table also includes the cumulative factor variances. The four identified latent variables collectively account for 48.9% of the variance within the dataset, with the variable PA0 contributing the largest portion of 19.1%.

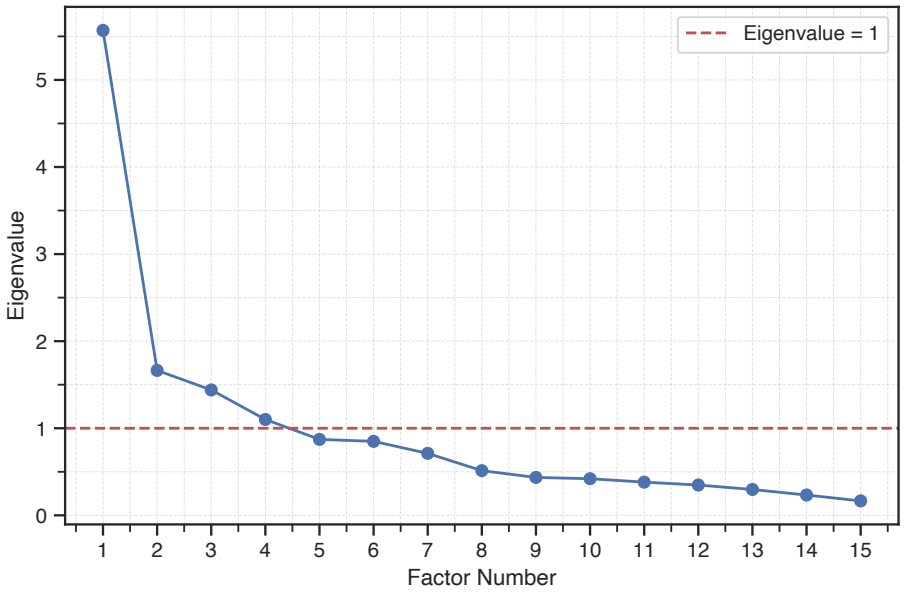


Figure 3.9: Principal Factor Analysis Scree Plot

The factor of greatest significance (PA0), which we term "Motivation and Perceived Ability" (MPA), is a latent variable measuring five of the MSLQ's motivation scales. These

Table 3.4: Principal Factor Loadings

Items	**Latent Variable Loadings**			
	PA0	**PA1**	**PA2**	**PA3**
Intrinsic Motivation	0.797	–	–	–
Learning Beliefs	0.759	–	–	–
Task Value	0.758	–	–	–
Self Efficacy	0.565	–	–	–
Extrinsic Motivation	0.563	–	–	–
Peer Learning	–	0.682	–	–
Critical Thinking	–	0.672	–	–
Help Seeking	–	0.646	–	–
Rehearsal	–	–	0.800	–
Organisation	–	–	0.660	–
Test Anxiety	–	–	–	0.531
Effort Regulation	–	–	–	0.521
	Cumulative Factor Variances			
	0.191	0.314	0.430	0.489

five scales loaded together on the factor and were all the scales that belonged to the Value and Expectancy components (Table 3.2). The Value scales, which include Intrinsic Motivation, Extrinsic Motivation, and Task Value, capture a student's desire to engage with the course and contribute the "Motivation" term to the construct's name. The Expectancy scales, namely Self-Efficacy and Control of Learning Beliefs, capture a student's confidence in their own ability and their belief that effort leads to success, and contribute the "Perceived Ability" term. Therefore, the MPA construct represents a student's innate desire to perform well in the course, combined with their own self-perceived ability to succeed in accomplishing this goal.

The second factor (PA1) is a latent variable that measures Critical Thinking, Peer Learning, and Help-Seeking. We term this latent variable: "Collaborative Cognition". This latent variable suggests that students who engage deeply in the course, such as exploring beyond the course material, often seek help and collaborate with peers. This seems consistent with the GNG1103 course design, relying as it does so heavily on teamwork.

The third factor (PA2) represents Rehearsal and Organisation, termed: "Plan & Review". This latent variable implies that students who distil the course material in their own terms, through notes and outlines, are more likely to go back and revise the material again. The final factor (PA3) measures Test Anxiety and Effort Regulation. We term this latent variable as "Anxiety-Driven Effort", representing the ability to study the course material under conditions which require more 'grit' [28] (e.g., when a course is boring or hard, etc.), measured as "test anxiety"-related fear in the course.

A path diagram showing the item (scale) loadings, along with the inter-factor correlations, is shown in Figure 3.10. Three inter-factor correlations are found above the 0.32 cut-off value as recommended in [77]. The presence of these correlations validates our previously chosen oblique rotation approach. The magnitude of these inter-factor correlations (ranging from 0.439 to 0.608) represent moderate relationships, suggesting distinct but related latent variables [21]. The latent variables PA0 and PA3 show the highest correlation at 0.608, indicating that students with high Motivation and Perceived Ability also exhibit elevated Anxiety-Driven Effort. Notably, the Plan & Review factor was the sole factor lacking correlation with any other factor.

An ongoing challenge in GNG1103, which is a first-year course, is creating a sense of student value for the planning process. Indeed, the current course objective is to teach students basic task management strategies (e.g. "doing what you say you will do" or "defining who is responsible for what") rather than having higher-level project management strategies, which are taught in a follow-on design course.

Among the 15 scales of Pintrich's Motivated Strategies for Learning Questionnaire (MSLQ), 12 are represented within our latent variables. The scales not included are Time and Study Environment, Meta-cognitive Self-Regulation, and Elaboration. This observa-

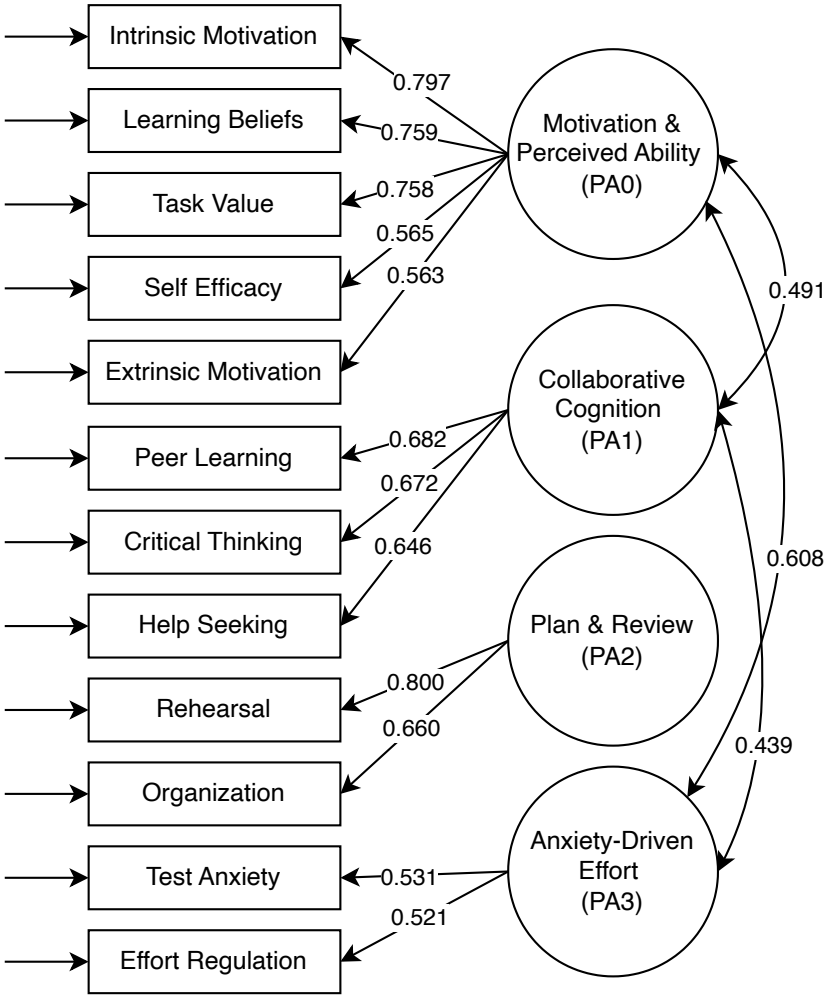


Figure 3.10: Path diagram depicting item loadings and correlation between latent variables

tion could be attributed to the nature of the questions in the Time and Study Environment scale. Pintrich scales were developed before the advent of "portable digital distraction infrastructure" such as smartphones. His questions focused on the locations where the study activities were conducted.

However, the significance of these locations might have diminished, given the pervasive and universal presence of smartphones and given the option of remote attendance or

asynchronous learning models. For example, Question 73, asks if they "attend class regularly". In the context of remote learning, a student may be present for the class yet not fully engaged with the course material, as they might be distracted by other tasks on their computer or smartphone.

Moreover, during the pandemic, many students occupied sub-optimal study environments as numerous households were constrained to share a single office space or forced to re-purpose bedrooms as study areas. Consequently, "attendance in the class" did not necessarily equate to cognitive engagement during the class. Indeed, it might not even have confirmed the presence of the student either. At the University of Ottawa, during the pandemic, it was decided that requesting a student to turn on their web camera, while potentially helping to improve student 'presence', constituted an invasion of privacy (i.e. forcing the student to invite the instructor and/or other students into their bedroom).

The absence of the Metacognitive Self-Regulation scale among the four most significant latent variables implies that reflection skills may not be an essential factor in the course. Since the study was conducted, a series of three reflection assignments have been designed and integrated into the course curriculum. These are designed to encourage students to reflect on what they have learned during the course, in the context of what they already knew beforehand.

Similarly, Pintrich's Elaboration scale comprises questions that are heavily associated with "lecture discussions" and "relating the course material to other courses that the student is taking". Since GNG1103 is experiential and project-based, the focus on lectures is reduced and designed to be synchronised with project activities or to required skills. In addition, the unique experiential nature of the course, in which students work with a client to tackle real-world projects, is unlike the majority of the other engineering courses at uOttawa, not including the design capstone and honours project courses.

Regarding the grouping of scales, among the six scales under the "Motivation" criteria of Pintrich's MSLQ, five scales were identified within our most significant latent variable (PA), "Motivation and Perceived Ability". The Test Anxiety scale was the notable exception. The Test Anxiety scale, on the contrary, is associated with the least significant observed

latent variable (PA3), namely, "Anxiety-Driven Effort". This result suggests that Test Anxiety may not be a significant factor in the minds of students for this design course. The 5 individually-assessed quizzes, each worth 3%, were distributed evenly throughout the course and the final exam was a summative test after course completion, worth 30%.

Among the nine scales that constitute the "Learning Strategies" criteria of Pintrich's MSLQ, three scales, specifically Elaboration, Metacognitive Self-Regulation, and Time and Study Environment, are notably absent from the identified latent variables. The remaining were loaded onto the less significant factors, namely, PA1 to PA3. This observation indicates that among these engineering undergraduate students, the Motivation scales of Pintrich's MSLQ questionnaire hold greater significance compared to the Learning Strategy scales. Based on this finding, MPA was used to measure learner motivation in the *Phooled!* study, where its influence in anti-phishing education was further examined.

However, among the scales that do not fall under the Motivation criteria, the Peer Learning scale had the most significant loading in our factors. This finding is surprising, considering the fact that this scale has the lowest average scores on all scales. This is because factor loadings reflect how strongly a scale relates to other scales, rather than the absolute values of the scale. This phenomenon may be attributed to the high emphasis on collaborative group activities within the course curriculum. The significance of Peer Learning motivated the inclusion of collaborative gameplay in *Phooled!*, where participants work together to detect phishing emails. A similar trend can be observed for the Help-Seeking scale, which had the third lowest average scores across the terms, but was the third most important scale outside of the Motivation criteria.

3.4 Conclusions and Recommendations

Overall, the MSLQ scales do not vary significantly across the three modes of instruction, namely: online, HyFlex, and in-person; with the notable exception of self-efficacy. Self-Efficacy for learning and performance in the course dropped significantly during the transition back to the in-person mode of instruction and was correlated with self-regulation.

Interestingly, Self-Efficacy values were significantly higher in the 2024 Winter term compared to the 2022 Fall transition term, suggesting that later in-person cohorts showed higher Self-Efficacy than the transition cohort. This increase in Self-Efficacy appears to place greater responsibility on the student rather than on the course designer.

Nevertheless, it appears that the "Affective" components, such as "Test Anxiety", warrant more careful consideration by the course designer. Generally, anxiety levels are higher in an engineering design course implementation, which is project-based, like GNG1103. This anxiety level also seems to vary widely between students. It would be interesting to investigate the sources of that variation (e.g., student motivation or ability, etc.) and to compare that with student performance, using assessments by peers or project managers or external judges and clients, that are already part of the course.

Most worryingly, peer learning and help-seeking scores were low in GNG1103, which would seem to be something that course redesign, or perhaps better mentoring, could address. Increasingly, the role of senior student project managers is valued by students, but the recruitment pipeline and "community" of these mentors was also dramatically affected and disrupted by the COVID-19 global pandemic. This pipeline of senior student members is being recharged, slowly, back to where it was in 2020.

In GNG1103, a 'job shadowing' strategy is used, where senior project managers mentor more junior ones, before being 'appointed'/paid. This process was partially broken by the pandemic, where sidebar or 'overheard' communications all but evaporated, between all students, including between the student mentors. The value of this kind of 'indirect' communication is probably worthy of further study, in the context of the teamwork in the design course and might be worth including in the training given to the project managers.

In our exploratory factor analysis, the latent variable "Motivation and Perceived Ability" emerged as the most significant, explaining approximately 20% of the dataset's variance. This finding indicates that the comprehensive MSLQ questionnaire, with 15 scales comprising 81 items, could be condensed into a five-scale questionnaire with 26 questions that effectively captures MPA. This could help to shorten the duration of the survey, thereby decreasing the likelihood of user survey fatigue and survey dropout rates. This

will be important for the *Phooled!* game design, where participants will have already spent time playing the game and extra time for the survey is less practical.

Moreover, our analysis suggests that the original questions provided by Pintrich for the "Time and Study Environment" scale may not adequately account for the realities of currently prevalent "digital distraction infrastructure", such as smartphones, particularly within the context of remote learning environments. We suggest that future research should aim to develop more appropriate questions to capture the influence of such distraction infrastructure. Further studies should be performed, with a prolonged online mode of instruction, to observe if the students' self-efficacy becomes comparable with that of the assumed (2024-level) in-person mode, on a longer-term basis. In general, it might be informative to compare or correlate these scales with overall performance in the course (e.g., using individual assessments like the final exam, or the peer-based assessment of personal group contribution. Equally, the overall final grade could be used, although this is highly dependent on the project design deliverable assessments). However, none of these were within the bounds of the ethical permissions that were obtained for this study.

The next chapter introduces *Phooled!*, an anti-phishing educational game designed to teach phishing avoidance behaviour. Several findings from this study directly informed the design of the *Phooled!* game. The exploratory factor analysis identified Motivation and Perceived Ability (MPA) as the dominant latent construct for measuring motivation in the MSLQ, accounting for 19.1% of the variance.

To study the effectiveness of this game, the MPA construct is used across both in-person and remote instruction modes, as detailed in Chapter 5. The concise MPA measure made it practical for use in a game-based study where participant time is limited. Since the motivation portion of the MSLQ includes Test Anxiety along with the subscales that load onto MPA, for ease of administration, the entire motivation section is used. Whether this motivation construct generalises beyond a course context to help predict learning in a specific domain, such as anti-phishing education, is an open question. The collaborative game elements in *Phooled!*, including collaborative phishing detection and real-time game chat, reflect the significance of Peer Learning identified in this study.

Chapter 4

Phooled!: Collaborative Anti-Phishing Educational Game Design and Implementation

The anti-phishing training approaches surveyed in Chapter 2 range from embedded phishing campaigns and in-person training to board games and simple digital games, yet each has limitations that motivate our choice to develop a remote and collaborative digital game. Embedded campaigns produce only marginal improvements, with little user engagement (Section 2.1.1). In-person training and board games require instructors or subject matter experts, who are in short supply, and therefore do not scale (Section 2.3). Existing digital games focus narrowly on URLs instead of the full email content, and are single-player rather than collaborative (Section 2.1.2).

Game-based learning has repeatedly outperformed passive learning material in improving phishing detection [72, 90]. It also engages learners through rich in-game interaction mechanics, such as reward incentives [57] and role-play [69, 90]. A remote digital game can scale as it does not require an expert to be present, and can include collaborative learning, which improves retention [46].

This rationale motivates the design of *Phooled!*, a collaborative anti-phishing training

game built on the in-game interaction mechanics and learning theory discussed in the preceding chapters. It was developed as an extension to Phishducation [1] an existing cybersecurity education web application at the University of Ottawa, and is modelled after traditional card and board games with an emphasis on remote learning. Section 4.1 describes the *Phooled!* game design, Section 4.2 details the experiment designed to evaluate the collaborative aspects and in-game interaction mechanics of *Phooled!*, Section 4.3 covers the *Phooled!* system architecture and implementation, and Section 4.4 details the email sample labelling process that serves as the game's sample database.

4.1 Phooled! Game Design

Phooled! translates the learning theory principles from Chapter 2 into in-game interaction mechanics. In this section, we describe the core aspects of *Phooled!*, while omitting any technical implementation details, which will be discussed in the next section. First, we walk through a complete game session in Section 4.1.1, then we describe how different learning theories were realised as in-game interaction mechanics in Section 4.1.2.

4.1.1 Game Rules and Walkthrough

Every *Phooled!* session consists of a battle between two or more players. The game has three main rounds, Quick Pick, Trick Pick, and Group Pick, followed by a Feedback and Review round, each described in their respective order below. Each round is designed to activate different message processing modes, as described later in Section 4.1.2. All in-game interaction mechanics below are described using their default durations, which can be modified before every game session. Figure 4.1 illustrates this game progression as a UML state machine diagram.

Phooled! is analogous to a card game, where each card is an email sample that is either Phishing or Safe. At the start of a game, $2n$ cards ($n = 3$ by default) are drawn without replacement from the email sample bank, balanced by prior exposure, then dealt

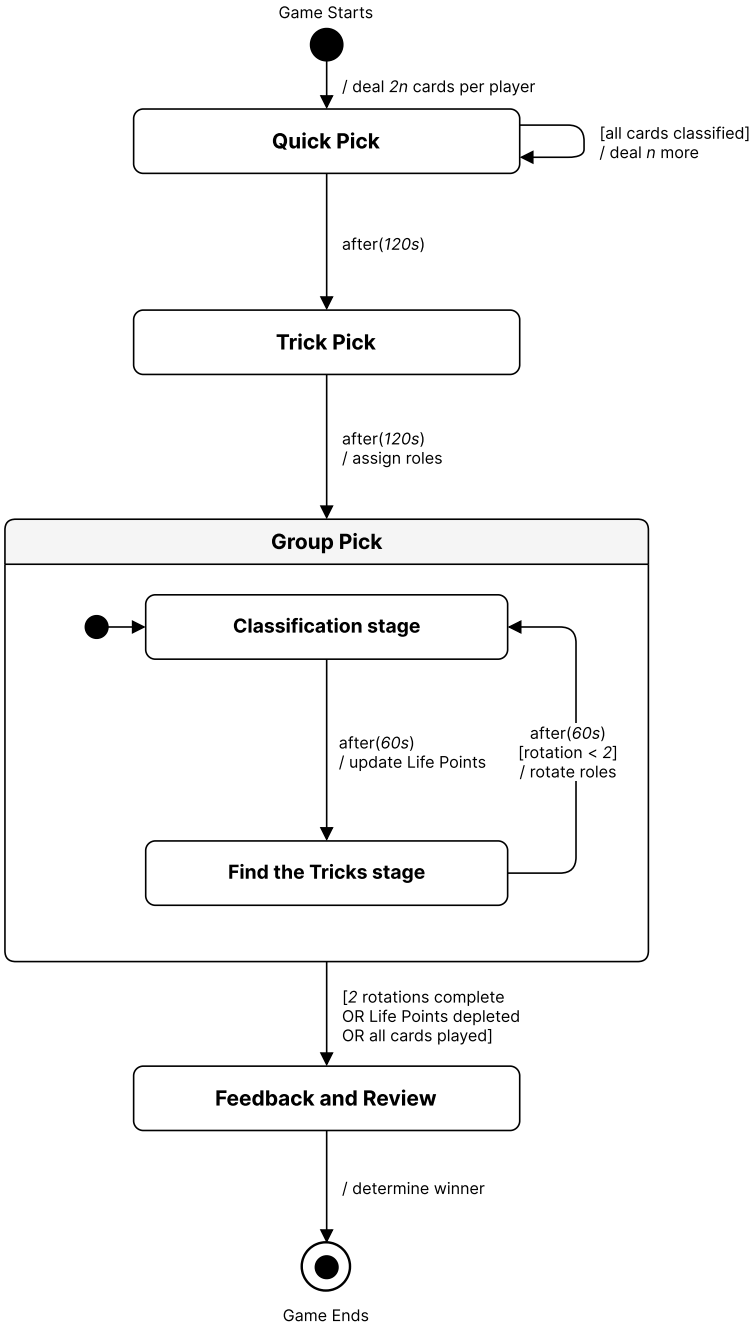


Figure 4.1: UML state machine diagram for a session of *Phooled!*

to each player. The objective of the game is to trick opponents into misclassifying emails while avoiding being tricked yourself and the game scoring is designed to reflect these objectives. Each player starts with two motivation tokens: three Life Points and zero Gold Coins. During the game, players lose Life Points or reduce their opponents' Life Points as described in the following rounds, but are ephemeral in a given game session. However, gold coins carry forward to different games in the future and are designed to increase longer-term engagement with the learning.

Quick Pick Round In this round, each player works individually to classify all their cards as either Phishing or Safe. By default, they have 120 seconds to do so. If they finish classifying all of the cards in their hand, an additional n cards are dealt to that player once. At the end of the allocated time, the Quick Pick round ends and the game proceeds to the next round.

It is important to note that players are not yet aware of whether their email classifications were correct. This is because, in the subsequent Trick Pick round, players are required to select cards to keep in their hand based on their perceived confidence in their classifications. Therefore, feedback about their classification correctness is not presented yet at this stage.

Trick Pick Round In Trick Pick, each player has 120 seconds to select n cards from their current hand of $2n$ to $3n$ cards by keeping some and discarding others. For example, a user's strategy might be to keep cards that fall in a sweet spot: ones the player is confident about, but that they believe will trick their opponent. At the end of the Trick Pick round, if the player has only selected m cards such that $m < n$, the remaining $n - m$ cards are chosen at random from their hand and the rest are discarded.

Group Pick Round - Classification Stage The next round is Group Pick, which consists of two stages: the "Classification stage" and the "Find the Tricks stage". When this round commences, each player is assigned a role as either: the Trickster, the Defender, or a Spectator. This round is played twice in one game, and in each rotation, the players

rotate roles in a round-robin fashion. First, in the Classification stage, the Trickster has 30 seconds (by default) to select a card from their hand to use as a phishing attack against the player designated as the Defender. The Defender and the Spectator wait for the Trickster's decision during this period.

Once the card is played, all players have 60 seconds to classify the card as either Phishing or Safe. Both the Defender and Spectator are encountering this email for the first time, but the Trickster has the advantage of having seen this card in Quick Pick and being confident enough not to discard it in the Trick Pick round. Once the time elapses, their classification answers are compared to the ground truth to determine whether the Trickster's phishing attempt was successful.

There are four possible outcomes, as shown in the truth table in Table 4.1, each with its own Life Point consequence. The classification results are then shown to all players. The Spectator's Life Points are never at stake in this stage. Instead, they receive a bonus of 50 Gold Coins if their classification is correct.

Table 4.1: Truth table of Life Point outcomes in the Classification stage

Trickster	**Defender**	**LP Lost By**	**Outcome**
Correct	Correct	–	Trick defended
Correct	Wrong	Defender	Defender tricked
Wrong	Correct	Trickster	Trick backfired
Wrong	Wrong	–	Both tricked

Group Pick Round - Find the Tricks stage In the next stage of the Group Pick Round, each player is shown a list of cues that could be present in an email, such as grammatical errors or link mismatches. The players have 60 seconds to select all the cues that are present in the email. Players can see each other's selections in real-time and communicate with each other during this stage. After the stage ends, each player's selections are compared to the ground truth.

A correct selection earns 5 Gold Coins, while each incorrect selection costs 5 Gold Coins. However, a player's Gold Coins cannot drop below zero. If an email contains no cues, players are rewarded with a bonus of 50 Gold Coins, as long as they have no false positive cue selections. The game ends when both rotations are complete, when any player's Life Points are depleted, or when all players' cards have been played.

Feedback and Review Round Once the game ends, a Feedback and Review round allows players to revisit each card that they had seen during the game. They can view their own classification and cue selection results, compare them with the correct answer, and raise disputes if they do not agree with it. Disputes that are approved by the game administrator receive bonus Gold Coins.[1] After the review, the player with the most Life Points is deemed the winner. In the event of a tie, the player with the most Gold Coins wins. Gold Coins that players obtained during the game carry over to subsequent game sessions. With the Feedback and Review round complete, the game session ends. The following section describes the in-game interaction mechanics that are embedded within these rounds.

4.1.2 In-Game Interaction Mechanics

The rounds in *Phooled!* map directly to the two processing modes of the Heuristic-Systematic Model, as described in Section 2.2. The Quick Pick round is designed to induce heuristic processing: a player must classify up to $3n$ cards (nine) within 120 seconds, and due to this time pressure, they are expected to default to their heuristic processing capabilities. Conversely, the Classification stage allocates 60 seconds to classify a single card, giving the player enough time to analyse the email's content and identify its phishing cues, thereby pushing them towards systematic processing. As described in Section 2.2, this simulates the real-world behaviour, when users are under time pressure, perhaps due to workload, such that they only switch to systematic processing after encountering a trigger.

[1] This also allows samples to be corrected, in cases where experts have not classified the samples properly or where there is ambiguity.

By rotating through roles, players are forced to think from the perspective of different actors involved in the phishing process. For example, a Trickster thinks from the perspective of a malicious attacker and considers the tricks they might use to deceive users. On the other hand, the Defender is placed in similar conditions to those of a real-world email reader.

However, when the group size of a game session is greater than two, we require the inclusion of an additional role called the Spectator. This role ensures that when the Trickster and Defender are battling, no player is kept idle, and that all players remain engaged in the learning activity. Since the Spectator does not have any Life Points at stake, they earn Gold Coins as a bonus for correct classifications instead.

In the Classification stage, Life Points, the primary driver of player behaviour within the game, are at stake. Inaccurate classifications can result in Life Point loss, which could eventually lead to elimination from the game. Thus, this mechanism ensures that players are motivated to classify accurately. As shown in Table 4.1, a player does not lose a Life Point when their opponent does not know the correct classification either. This ensures that skilled players cannot be eliminated by pure chance.

The game has Gold Coins as a secondary incentive, primarily to encourage collaboration in the group stage. By having a negative Gold Coin penalty, we force the player to systematically process the email and identify only the correct cues, rather than selecting all of them. The 50 Gold Coins bonus further discourages false positives by rewarding correctly identifying emails with no phishing cues.

The Phishducation platform carries over Gold Coins to subsequent *Phooled!* game sessions, providing a long-term incentive for players to return and keep learning anti-phishing behaviour. The Gold Coins gained per attribute are scaled by a group multiplier factor that depends on the number of players that agree on that attribute, which encourages collaborative play rather than competition among the players. Both Life Points and Gold Coins can be described as extrinsic motivators (Section 2.4) for short-term and long-term play respectively.

Feedback on their performance is provided to players at two points, inside the game itself

and at the end, as part of the Feedback and Review round. The in-game feedback occurs at the end of the Classification and Find the Tricks stages. For the classification feedback, the ground truth of Phishing or Safe is shown along with the player's own classification correctness. For the Find the Tricks stage, all the ground truth cues are highlighted in their actual positions in the email. And finally, in the Feedback and Review round, both classification and attribute feedback are provided together for all cards encountered by the user during the game.

The in-game feedback fulfils the principle established in Section 2.3, that feedback should be close to the learning activity and connected to it. Similarly, all types of feedback are presented as part of the game, instead of being presented as a secondary or post-learning activity. The dispute mechanism further encourages the player to critically examine their feedback and raise disputes about the sample bank ground truths.

The previous sections describe the default Collaborative game mode. However, for experimental purposes, as described in Section 4.2, we modify the game to have a Single-Player version. In this version, the player alternates between the Trickster and Defender roles, and a coin toss decides the classification outcome of the opponent role. Similar coin tosses are used to decide which cards to keep or discard in the Trick Pick round, and which card to play as the Trickster. The Find the Tricks stage is played individually in this variation.

4.2 Comparative Study of Individual versus Collaborative Gameplay for Anti-Phishing Education Effectiveness

In Section 2.7, we formulated four RQs, based on our literature review of existing methods in the anti-phishing training domain. To address these questions, we designed an experiment to evaluate *Phooled!* using Pre-Test and Post-Test assessments, as well as in-game performance. This experiment compared three treatment conditions: Control,

Single-Player, and Collaborative treatments, across three different cohorts.

4.2.1 Research Questions and Hypotheses

We now present specific hypotheses for the confirmatory research questions RQ1 and RQ2, and outline the scope of the exploratory research questions RQ3 and RQ4.

RQ1: Phishing Detection Improvement.

H 1. *Training with Phooled! improves phishing detection performance, relative to the Control group.*

H 2. *Collaborative training outperforms Single-Player training.*

RQ2: Predictors of Phishing Detection Improvement.

H 3. *Subjective norms influence phishing detection improvement.*

H 4. *MPA influences phishing detection improvement.*

We also explore whether instruction mode (remote vs. in-person) affects phishing detection improvement.

RQ3: Evaluation of In-Game Interaction Mechanics. Using our in-game user interaction data, we explore the effects of our in-game interaction mechanics on phishing detection performance as well as improvement in the detection ability of our research participants/learners. Specifically, we look at the effect of collaborative cue detection on knowledge transfer, and whether elements such as the player roles, feedback and review, and in-game chat improve learning and help keep learners engaged.

RQ4: Determinants of Phishing Email Detection Difficulty. We examine classification and cue detection performance inside *Phooled!* to analyse how email features, such as sample type, phishing cues, and sender details, influence the difficulty of detecting phishing attempts.

4.2.2 Experimental Procedure Stages

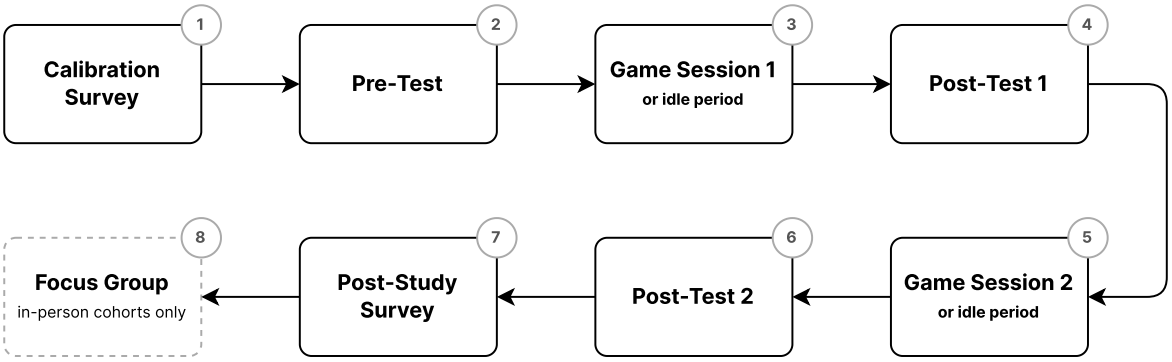


Figure 4.2: Flowchart of the *Phooled!* experimental procedure

Our experimental procedure, as illustrated in Figure 4.2, consisted of two game sessions, each followed by an assessment. The type of game session, whether Single-Player or Collaborative, was determined by the participant's treatment assignment. Prior to any training, baseline phishing detection ability was assessed using the Pre-Test. Additionally, a calibration survey at the outset measured participants' individual characteristics using our survey instruments.

At the end of the experiment, participants completed a post-study survey, along with a focus group for in-person cohorts. All surveys were administered as embedded Google Forms within the Phishducation platform, with a sequential checkpoint system gating each game session until the prerequisite surveys were completed. Participants were incentivised with 50 bonus motivation tokens for each survey completed.

This study used a mixed factorial design, with each phishing detection assessment (i.e., Pre-Test, Post-Test 1, Post-Test 2) as the within-subjects factor, and treatment condition (i.e., Single-Player, Collaborative, Control) and cohort (i.e., IT Professionals, HS Students, Remote) as the between-subjects factors. Three performance measures were obtained from these assessments: Test Scores (classification accuracy score out of 10), Test Duration (total test duration), and IES [82, 83], which captures the speed-accuracy tradeoff, calculated as IES = Duration/Score.

Participants were randomly assigned to one of three treatment conditions:

1. **Single-Player**: participants played *Phooled!* against PhishBot.
2. **Collaborative**: participants played *Phooled!* in groups of 2–5.
3. **Control**: did not play *Phooled!*, observed an idle period instead.

All features of the game except the collaborative component were kept constant across the Single-Player and Collaborative treatments. Though our main goal was to study the effectiveness of Collaborative training compared to Single-Player training, it was necessary to include a Control group to first establish whether the game itself had any effect on phishing detection improvement. Participants in the Control group followed a 20-minute idle period between each assessment. This duration was calculated from the average game session duration in the in-person cohort part of the study.

The sample sizes of the two in-person cohorts were relatively small ($n = 14$, $n = 15$), so the inclusion of a Control group would have left too few participants per treatment for meaningful comparisons. Additionally, practical constraints of conducting research in schools and organisations as a learning activity meant that all participants in those cohorts had to interact with learning material. The Remote cohort ($n = 35$) was larger than both in-person cohorts combined, which allowed for the inclusion of a proper Control group who had no interaction with training material. The IT Professionals cohort completed only one game session due to time limitations of the in-person session. Therefore, Post-Test 2 data in this study is constrained to HS Students and Remote participants.

4.2.3 Study Participants and Recruitment

Participants were recruited from two in-person cohorts and one Remote cohort that participated synchronously online. The HS Students cohort was recruited from the Faculty of Engineering Secondary School (FESS) at the University of Ottawa, and the IT Professionals cohort was recruited from the uOttawa Cybersecurity Office. An additional effort

was made to target the general university population but was not pursued further due to limited turnout. The absence of compensation or rewards likely contributed to this low turnout. The experiment was run for the in-person cohorts in planned 3-hour sessions at the uOttawa campus.

Exam conditions were observed, where participants could only communicate with the researcher, or with other participants only within the game itself in the Collaborative treatment. The in-person participants either used the desktop computers present at the uOttawa CyberRange or their personal laptops to take part in the study. Participation was voluntary for the in-person cohorts, though they were provided with refreshments. Participants recruited from the Prolific platform were compensated $25 for taking part in the study. The full completion of both game sessions was required for inclusion in the data analysis. Ethics approval for this study was obtained from the University of Ottawa Research Ethics Board (REB), under the file number: H-02-24-10020.

4.2.4 Measurement Instruments for Motivation and Phishing Susceptibility

This section describes the survey instruments, along with the quantitative and qualitative assessments used to collect data in this study. The survey instruments are enumerated in detail in Appendix A. A calibration survey administered before the intervention measured six constructs: heuristic processing and systematic processing (SCAM scale), two subjective norms scales, Motivation and Perceived Ability (using the MSLQ), and Test Anxiety (MSLQ). The MPA scale was derived from the MSLQ analysis in Chapter 3, where it was identified as the dominant motivational construct, explaining approximately 20% of the variance in the collected data. Table 4.2 summarises the instrument.

The heuristic and systematic processing scales were adapted from the SCAM [86]. Two slightly different versions of subjective norms were measured. Subjective Norms 1 (SN1) [42] is a measure of general peer influence, and Subjective Norms 2 (SN2) [71] is primarily concerned with social influence from authority. The MPA as described in

Table 4.2: Calibration survey instrument summary

Scale	Source	Items	Response Format
Heuristic Processing	[86]	3	5-point Likert
Systematic Processing	[86]	3	5-point Likert
Subjective Norms 1	[42]	3	5-point Likert
Subjective Norms 2	[71]	4	7-point Likert
Motivation and Perceived Ability	[64]	26	7-point Likert
Test Anxiety	[64]	5	7-point Likert

Chapter 3 was measured using a reduced subset of the MSLQ [64], along with Test Anxiety from the same questionnaire.

Each participant completed three phishing detection assessments during the study: one Pre-Test and two Post-Tests, delivered after each game session respectively. Each assessment consisted of 10 email samples, evenly split between Phishing and Safe samples. All samples were selected from our email dataset and were unique to each assessment, as they were removed from the *Phooled!* sample pool. The difficulty of each test was calibrated with six medium and two each of easy and hard samples, as described in Section 4.4.

The email samples within each assessment were constant, but the order of the assessments themselves was randomised. Participant classification responses and test durations were recorded for each assessment. The assessments were based solely on email classification, mirroring the real-world email reading task. This is unlike the game itself, which also focuses on cue identification, a subtask of classification itself. We use performance metrics derived from these test scores and duration as the dependent variables analysed in Chapter 5. Within *Phooled!*, in-game user interaction data was logged, including classification decisions, duration per classification, and in-game interaction mechanics engagement.

A post-study feedback survey was administered after the final assessment. The survey covered the themes of their experience using *Phooled!*, their gameplay preferences, and their perceptions about software usability, engagement, and educational value. Most questions

used a Likert scale, while others were open-ended. For in-person cohorts, an interactive debriefing and focus group session was held for approximately 30 minutes at the end of the experiment. The focus group was semi-structured and focused on topics such as their *Phooled!* gameplay experience, their opinions on collaboration versus competition, and their prior exposure to phishing. Field notes were taken by the principal investigator while the focus group was conducted, and are discussed in detail during the qualitative data thematic analysis in Section 5.5.2.

4.3 Detailed Design and Implementation of the Phooled! Game

In this section, we describe the implementation of the *Phooled!* game design presented in Section 4.1 as an online real-time collaborative anti-phishing game playable within a web browser on a laptop or desktop computer, ensuring wider reach to users. We begin with the game user interface design in Section 4.3.1, then detail the system architecture that drives *Phooled!* in Section 4.3.2. The core design constraint was real-time coordination of game clients and synchronisation of server and client states, which influenced the system design decisions described in this section.

4.3.1 User Interface Design

We first present the game user interface as implemented, walking through the main user flow and explaining how some key in-game interaction mechanics are presented in the UI.

Game Setup and Coordination Interface The Active Games page lists all active game sessions. From here, players can either join an existing game session that is looking for more players, or host their own game. If a player decides to host a new game, a uniquely identifiable Game Room is created and an easily shareable link is generated that the host can send to other players to join.

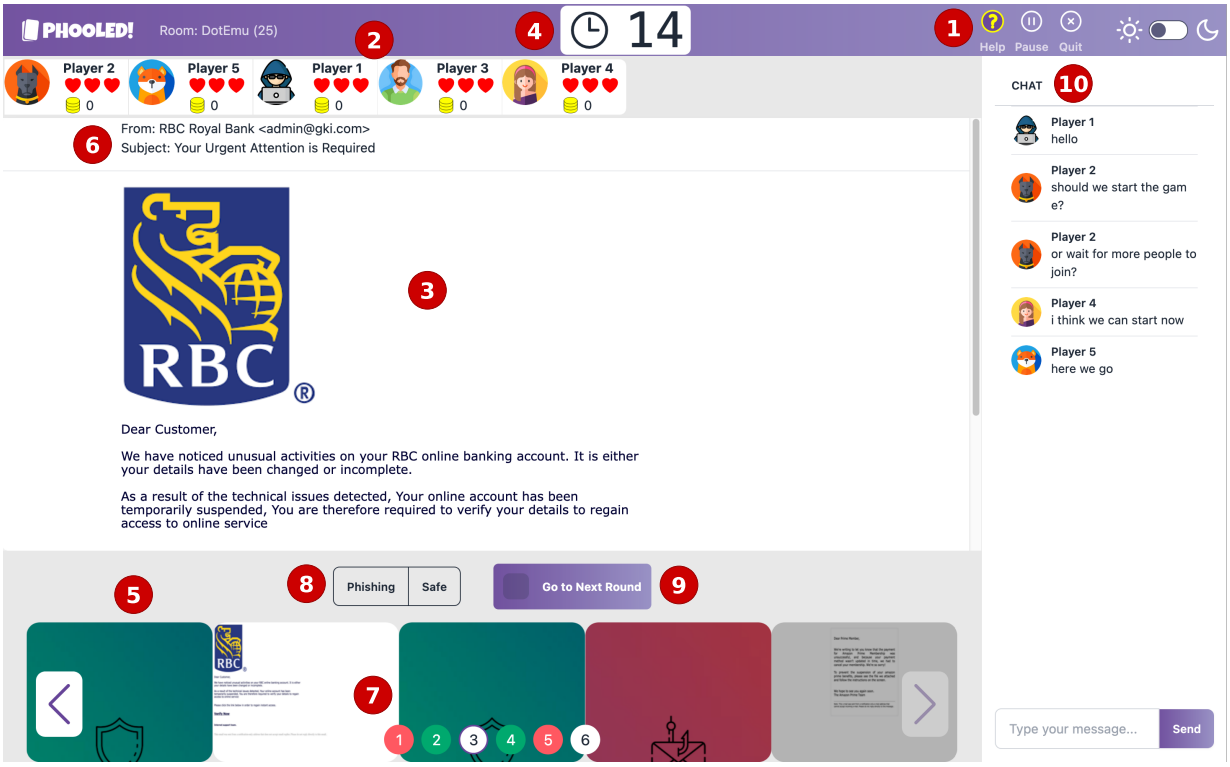


Figure 4.3: Email Classification task during Quick Pick round of *Phooled!*, with key elements highlighted

Once players join a game, they enter a Game Lobby where they can see other players and their avatars (i.e., profile pictures), chat with them in an in-built chat area that is available throughout the game duration, and signal that they are ready to play by using the Ready button. In the lobby, the host can configure game settings such as the durations of each round, and the number of cards (i.e., n) that will be used in Group Pick. Once all players have confirmed their ready status, the host can start the game, kicking off the Quick Pick round.

Game Instructions and Help Interface Throughout the game, instructions about what to do and general game rules are presented to the player as and when required. This is through two mechanisms: first, through "info screens" before every round with the rules

for that round. The round only commences once players have read the rules and readied up. The info screen has a visual timeline depicting the progress of the game, showing completed parts of the game and indicating the current status in relation to the end of the game. Furthermore, at any point during an active round, players can click the Help icon (marked (1) in Figure 4.3) on the navbar at the top of the screen. This lets them review the same instructions at any point they are stuck, however, this will not stop the timer from ticking down.

A player information area (marked (2) in Figure 4.3) at the top of the screen, above the email display area (3), displays all current players with their avatars, i.e., profile pictures, usernames and current Life Points (depicted as hearts) and Gold Coins. Whenever a role is assigned to the player, this is also indicated with a badge below their player information. The top navigation bar also displays a large countdown timer (4) for the current round, an indicator for the current progress of the round rotations, and buttons to pause or quit the game (1).

Quick Pick Round Interface

In the Quick Pick round, the cards in a player's hand are visualised in a card carousel interface (5), as shown at the bottom of the page in Figure 4.3. Each card contains an email from the labelled sample bank. The labelling procedure of these email samples is described in Section 4.4. The player can browse through each card by clicking on "face down" cards in their hand, turning them "face up", which then displays the full content of the email in the main email display area at the centre of the screen. The email is rendered using its actual HTML code, preserving its original appearance.

This means that the text content, images, and URLs appear no different from how they would look in a real-world email client like Outlook or Gmail. The email display area is scrollable to accommodate longer emails that extend beyond the height of the user's browser viewport. All links are shown with their original URL target destination, but have been programmatically defanged so that clicking on them does not navigate the browser away from the game to these malicious websites. The original email header information,

including the details of the sender, i.e., the "From Address" and "From Name", along with the Subject field, is displayed above the email content area (6).

Both the displayed email headers and content are personalised for the player. This is done by replacing the actual information of the original email receiver with those of the player, using the email address and username tied to their Phishducation platform account. The recipient details in each sample were replaced with placeholders as described in email curation (Section 4.4), and the game client substitutes these placeholders with the player's details when the email is rendered. For example, an email that originally included the text "the account tied to your email: `victim@example.com` has been locked!" would instead read as "the account tied to your email: `john123@uottawa.ca` has been locked" if the player had signed up using their university email account. At the top of the screen, a countdown timer is prominently displayed, showing the time left in the round. This timer is a key mechanism for inducing heuristic processing, as discussed in Section 4.1.2.

Below the carousel, numbered pagination circles (7) indicate the currently selected card, and clicking on the nth circle jumps to the corresponding nth card in the player's hand, in both the carousel and the main email display area. When a card has been classified, it is indicated in the carousel with a phishing or safe card image instead of the generic face-down card with the *Phooled!* logo. The pagination circles also reflect the classification, changing colour to indicate whether the card was classified as Phishing or Safe, i.e., red or green respectively. Within the game, distinct icons (e.g., a shield for Safe or a fishing hook for Phishing) and text labels are used alongside colour coding to make the game accessible to colour-blind users.

At the bottom of the screen, there are buttons to classify each card as either phishing or safe (8). A "Go to Next Round" checkbox (9) allows the player to ready up for the next round. In each round, the checkbox is disabled until the main objective for that specific round is satisfied. For example, in Quick Pick, it is enabled only after the player has classified all cards in their hand. The game advances to the next round when either all players have readied up or the round timer expires. A chat sidebar (10) is available throughout the game for players to communicate. The chat data collected during the study is analysed in Section 5.3.4.

Trick Pick Round Interface In the Trick Pick round, we replace the classification decision buttons with Keep or Discard buttons. When cards are discarded, they are greyed out or dimmed in the carousel as a visual indication of their state. The pagination dots of the carousel also reflect this, turning grey for discarded cards. In this round, readying up for the next round is only possible once n cards have been "kept" in the user's hand.

Group Pick Round Interface The game UI is adapted for each role to mirror the role responsibilities described in Section 4.1.1. For example, while the Trickster picks a card, other players see a waiting screen. Once the Trickster plays a card, it is displayed to all players in the game in the main email display area and classification occurs in the same manner as in the Quick Pick round.

After the stage ends, a feedback screen is shown, which prominently displays their classification correctness. The feedback screen shows the player's classification alongside the ground truth, and the result of the battle, i.e., whether the Defender got tricked or not. The loss of Life Points from the battle is shown using a message and is also reflected in the player information area.

In the Find the Tricks stage, the classification decision buttons are replaced with a grid of attributes (marked (1) in Figure 4.4) with each having its own checkbox (2). The name of each attribute is displayed alongside an information icon (3) that can be hovered over for a tooltip briefly describing that particular attribute. As other players select their attributes, their selections appear next to the attribute in real-time.

This is indicated with an overlapping circle of the group's selections with their avatar and a hover action with their username (4). After this stage ends, the total Gold Coins gained as a result of the player's attribute identification is shown (marked (2) in Figure 4.5), and if an email has no phishing cues, a special message indicates this. Feedback is also provided through colour-coded badges (3) indicating whether each phishing cue was correctly identified, missed, or selected in error.

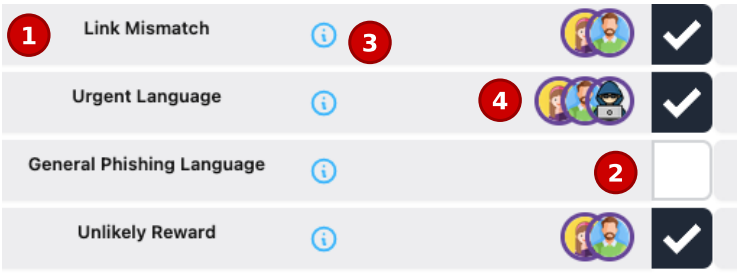


Figure 4.4: Collaborative Phishing Cue Identification in *Phooled!*, with key elements highlighted

Feedback and Review Interface In the Feedback and Review round, the card carousel is populated with all the emails encountered by the player throughout the course of the game. The classification and attribute feedback from each Group Pick stage are presented together for each card. The player can also raise their disputes about the ground truth of both types of feedback. A "Dispute Feedback" button allows the player to submit their opinion about the classification correctness of the card and a similar dispute can be raised for attributes through an attribute feedback modal. Both types of disputes are text-based and incentivised with a reward of bonus Gold Coins.

There is no set time limit for this round, which allows players to review their feedback at their own pace. This design choice was made to encourage systematic and deliberate processing of the feedback, with the absence of any time pressure. Once all players ready up for the next round, the final screen of the game with a leaderboard showing the winner is shown. It lists all players with their Life Points and Gold Coins, with a trophy for the winner.

4.3.2 Real-time Client-Server Synchronization Issues

In Section 4.3.1, we saw the design of the user interface for our game, and the rationale behind its various parts. In this section, we look at the design and implementation of a backend server that can drive our game logic and coordinate our clients' remote endpoints. Then, we briefly examine how the user interface, i.e., the *Phooled!* Frontend Client, is

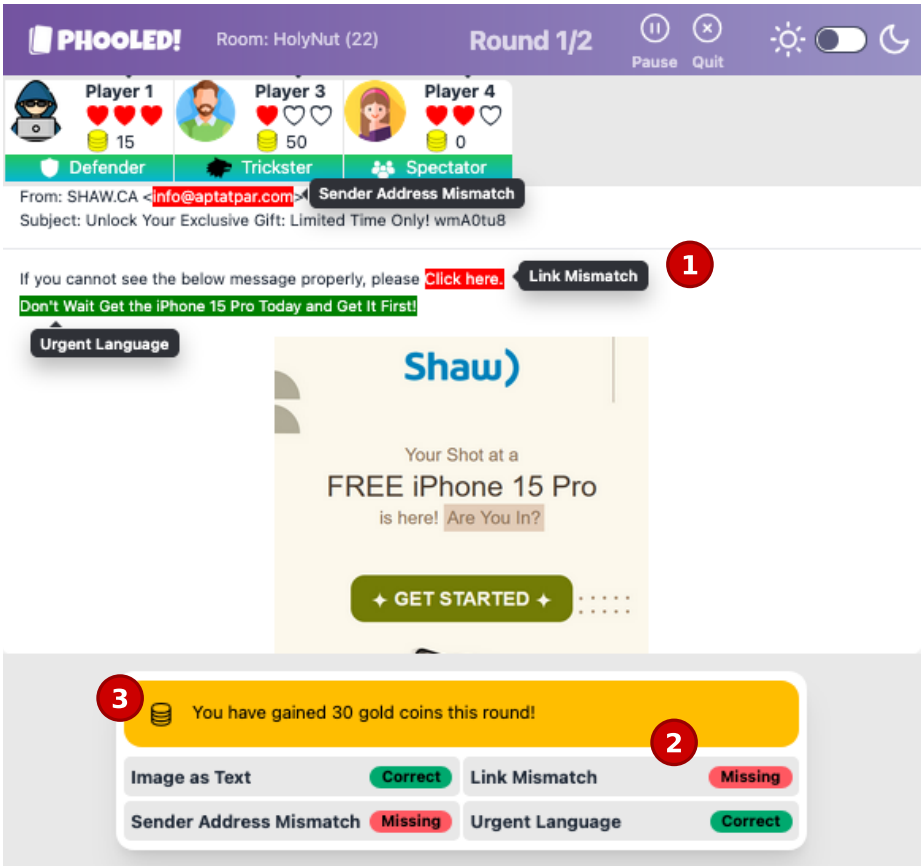


Figure 4.5: Feedback for Cue Identification in the Find the Tricks stage of *Phooled!*, with key elements highlighted

implemented.

The Phishducation platform allows users to create an account to watch anti-phishing training course videos, take phishing detection quizzes, or play a simple game called BetaPhish. *Phooled!* was developed as a new extension to this platform, largely because it already provides user management, database access through MySQL, user session tracking, and a variety of other functionalities that put the focus on the development of the game itself.

This existing Phishducation platform is built on Django, a Python web framework that implements a Model-Template-View (MTV) architecture, as opposed to the traditional

Model-View-Controller (MVC). In MTV, the models define the data layer, templates handle presentation (the view component in MVC), and views manage request routing and logic (the controller component in MVC). It is backed by a MySQL database for persistent storage of model information, like user accounts and user sessions.

The Django web application output is minimal, consisting of static HTML, CSS, and JavaScript-based web pages. The content from the website, both the app itself and static resources, is served through an Nginx web server. The platform is hosted on the Compute Canada cloud server. Therefore, this existing infrastructure was a design constraint on our design choices in this system.

In addition to this, three key design constraints guided this system architecture:

1. **Real-time bidirectional communication.** Inside the game, we require events like classifying cards, sending chat messages, progressing to next game states, etc. to be propagated between clients, i.e., the clients must be able to "talk" to each other during the course of the game. This requires a bridge or central connection point between these clients, which is the server in a client-server architecture. Specifically, a low latency, real-time duplex connection between each individual client and the server is required. In this way, the clients can talk to each other, using the server as a central relay node.

2. **Client-server state synchronisation.** Not only does this multiplayer game require communication between the clients, the state of each client needs to be synchronised. For example, even for the simplest base case of changing between game rounds, this change needs to happen together for all clients at the same time. Other actions, like the Trickster choosing a card to play in Group Pick, needs to trigger that card's email contents to be shown in other clients as well. Essentially, we need to maintain a global game state, allow clients to modify this global state through set game actions, and propagate these state changes to all clients in the session such that they are all synchronised.

3. **Scalable concurrency.** Our system must be able to support two types of concur-

rency. First, within a single game session, multiple clients may send actions to the server simultaneously, and these must be processed without conflicts or corruption of the game state. Second, multiple game sessions must be able to run independently at the same time, each with their own isolated game state.

The following subsections describe how each of these constraints was addressed in *Phooled!*.

Solving Client-Server Communication Constraints using WebSockets

The request-response paradigm of a standard Django application needs to be augmented to allow for this persistent bidirectional connection that exists throughout the game duration. We elected to go with a well-established WebSocket approach. This approach allows us to maintain a real-time connection between each client and the server under the constraints of low latency and low bandwidth.

We implement this using the Django Channels third-party library, which adds capability for creating, managing, and communicating using WebSocket connections. We use Daphne, an Asynchronous Server Gateway Interface (ASGI) server, as a router to direct WebSocket connections to Django Channels, and standard HTTP requests to Django's internal routing handler. The server maintains a uniquely identifiable WebSocket connection for each player.

Each such connection is served by its own thread of the WebSocket (WS) Consumer component, as illustrated in the system architecture diagram in Figure 4.6. These connections are bound in a unique WebSocket Channel corresponding to individual game sessions. The WS Consumer can receive and send messages from each client, and can use WebSocket's broadcast feature, within a channel, to efficiently replicate game actions across all connected clients in a game.

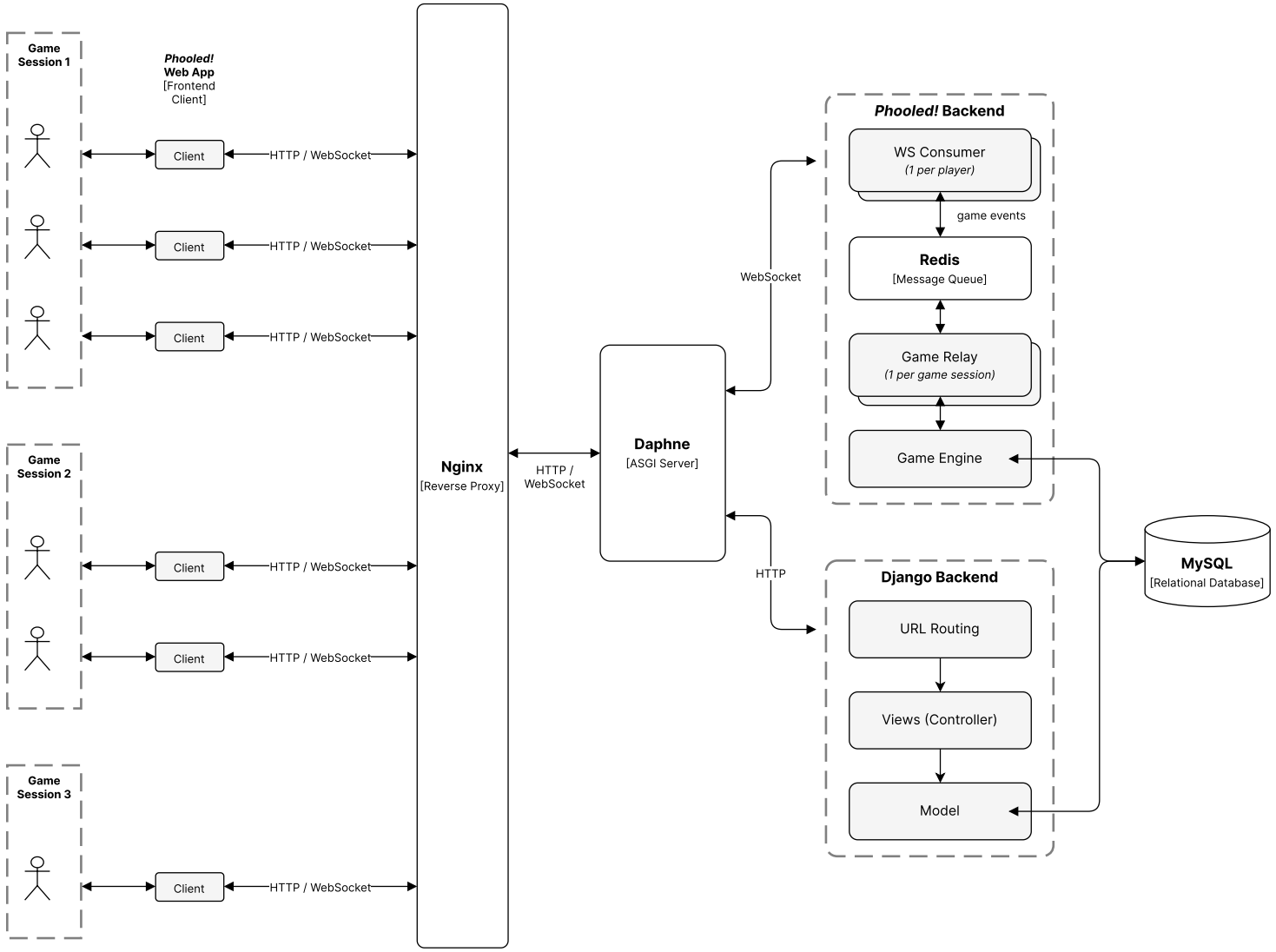


Figure 4.6: System architecture diagram for *Phooled!*, with software components developed for this study shaded in grey

Game Engine Design to Drive Game Logic using FSMs

We developed the Game Engine as a component responsible for maintaining game state and applying game logic. It is powered by a Finite State Machine (FSM) representation through the use of the `fsmpy` Python library. The actual FSM implemented is in fact a more granular version of the FSM depicted in Figure 4.1, with intermediary transition states in practice. Thus, any current game state is always bound to be valid, otherwise the transition would have been limited by the FSM specification.

The Game Engine is then the arbiter of all game state, every game action from a client is sent to the Game Engine through the use of a relay process that calls the Game Engine library. There is one server-side thread that maintains a Game Engine for each currently running game. This ensures that multiple concurrent game sessions can function in parallel without affecting each other's state.

The Game Engine processes game actions by following game logic to change state and update the persistent database. For example, the last player's "ready" action triggers a progression of game state while the "classify" action checks the database for ground truth. The Game Engine also takes care of state progression by making asynchronous state transition function calls that expire when the round is set to expire.

Each round is associated with a configurable time limit that can be set by the host during the game lobby state. If a client does not perform an action within the time limit, the Game Engine takes an arbitrary action to progress the game. For example, the Trickster not selecting a card to play within 30 seconds forces the Game Engine to forcefully select a random card from their hand. If a player's connection is lost, the game ensures resiliency by detecting the loss of a WS connection and transitioning the Game Engine into a Pause state, which is present in the extended FSM.

This allows the client to rejoin the game and re-sync its state with the Game Engine to continue. The game can be restarted without the player rejoining by kicking the player out of the room, ensuring that the de-synced client can never rejoin the session again. If all players disconnect, the Game Engine waits a set timeout period (10 minutes), at the end of which it marks the game as abandoned in the database and the thread is killed.

Client-Server Messaging Protocol

The client and server needed a common protocol to communicate with each other over these WebSocket connections. We propose a simple but efficient messaging protocol over this network. The protocol was encoded using JSON and specified two parts to each message: a message code, which is an integer that maps to specific game actions, and a payload, which is the data that is required to perform that action. Each type of node in the system: the server, the host, or a player, had specific actions assigned to it that it could communicate with under this messaging system. The detailed implementation of this custom message protocol can be found in Appendix B.

In-game chat communication bypassed the overhead of the Game Engine. Chat messages were directly broadcast to all other clients instantly by the WS Consumer without any additional processing. The execution of all game actions was done in a sequential manner. Actions from the client were queued as they were received by the WS Consumer and stored in a separate Redis message queue. The Game Relay thread polled a blocking pop operation with a minimal timeout on this queue such that it repeatedly "woke up" to process these incoming messages. Server communication was sent directly via the Game Engine using asynchronous calls.

The Game Relay was run as an asynchronous runtime managed by the `trio` library. This design choice was made as this library guaranteed that each game action was executed through the Game Engine strictly one at a time, thus transforming asynchronous aspects of the execution into a sequential flow when message processing order mattered. This feature, along with the First In, First Out (FIFO) nature of the Redis queue, ensured that race conditions were avoided by virtue of the system design alone without the need of a explicit locking mechanism like mutexes.

Frontend Client Design to Implement User Interface

The interactive and real-time nature of *Phooled!* requires the frontend client to dynamically update its UI in response to changes in game states without reloading the web page.

The existing static Django implementation can only support page updates with reloads. The current industry standard would be to use a single-page application (SPA) framework such as React. However, due to the high effort in rewriting the existing platform backend, along with the benefit of Django's built-in infrastructure for user and data management, we decide to minimally extend Django's capabilities using a plug-and-play approach with Alpine.js. Alpine.js is a minimal library that adds reactivity to static HTML pages by injecting a JavaScript script, thus extending the capability of Django's static HTML templates.

The UI is built using the DaisyUI framework which leverages TailwindCSS components as the foundation. The card carousel functionality is built using the Swiper.js library. We handle the WebSocket messaging natively through the browser's JavaScript implementation. There is no FSM on the client side, only reactive asynchronous functions that listen for game states and trigger UI updates, ensuring that the server always remains as the single source of truth. The entire frontend client is bundled using Webpack and served using the built-in Django web server.

Single-Player Implementation As introduced in Section 4.1.2, for experimental purposes, we develop a Single-Player mode of *Phooled!* with an automated opponent. For this game mode, the Game Engine spawns a PhishBot opponent that performs basic automated behaviour. It appears in the player information area as an additional player, and game behaviour and instructions are modified to support Single-Player gameplay. PhishBot performs game actions according to the logic described in Section 4.1.2, with the implementation using random functions in place of the theoretical coin tosses.

4.3.3 Collection of In-Game User Interaction Data

Phooled! includes many game interactions that could be studied to model user phishing susceptibility and detection behaviour. As such, these rich in-game interaction mechanics are useful for exploratory study as they can generate log data from user interactions, commonly referred to as telemetry data, as reviewed in the literature (Section 2.1.3).

Every event within this client-server system is captured as timestamped telemetry data in a persistent database, along with some additional events to track user behaviour.

For example, a telemetry event is sent every time a card is opened in the in-game display area such that when the card classification event is received by the server, we can calculate the amount of time spent classifying that particular card. Similarly, player card viewing behaviour and cue selection behaviour is tracked. Thus, these stored telemetry events are used to explore user phishing detection behaviour in Chapter 5. Additionally, the analysis of such data can help us validate our assumptions about our in-game interaction mechanics, and whether they are beneficial to learners.

4.3.4 Testing and Quality Assurance

The relative complexity of the system required diligent testing of game states and client-server interactions. The testing strategy aimed to target both the frontend and the backend through unit tests and integration tests to monitor game crashes, corrupt game states, and response to random and unexpected client actions. The FSM and game logic were validated through unit tests written using `pytest`.

It also tested the custom message protocol by mocking game actions and validating the subsequent state changes. The integration tests performed end-to-end game simulations by executing five browser instances in parallel using Selenium and Mocha.js. These clients played using randomised behaviour to progress until the end of the game with no game crashes or UI glitches.

The entire infrastructure, both the backend and the testing, was run on Docker containers for reproducibility in environments, such that local and cloud behaviour was guaranteed to be identical. All development work ran on a Continuous Integration/Continuous Deployment (CI/CD) pipeline that validated each code commit against both types of tests. These updates were then pulled into the cloud server and their Docker containers restarted with validated commits. Multiple internal betas were conducted and a pilot study was executed with around 10 individuals. The feedback and bugs captured during this process were iteratively fixed within the game before launch of the study.

4.4 Process for Email Classification as Phishing or Safe and Identification of Phishing Cues

This section details the creation of a labelled sample bank of real-world emails. These will be fed into our platform to power the *Phooled!* game, serving analogously as the "cards" in the game. First, we describe the data labelling pipeline and consensus-based annotation process in Section 4.4.1. Then, we present the cues that were marked in our email dataset in Section 4.4.2.

4.4.1 Semi-Automated Email Classification and Cue Identification Processes

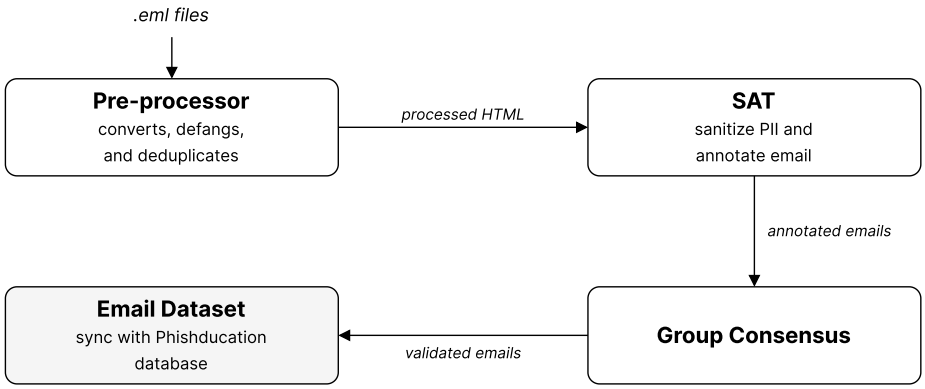


Figure 4.7: Flowchart of email dataset labelling

One of the main areas of concern facing anti-phishing education is the need to train users on the latest threats which are always adapting over time, as detailed in Chapter 2. Therefore, we need a scalable solution that helps us easily update our email dataset. To this end, we present a semi-automated process to transform ordinary emails from clients such as Outlook into email samples that are compatible with the full range of interactivity as described in Section 4.3.1. These were sourced from real emails delivered to university

email inboxes and exported as .eml files. The email dataset creation process is illustrated in the flowchart shown in Figure 4.7.

A pre-processing script first compares a hash of the content of each exported .eml file against the existing email dataset, discarding duplicates. It then converts the remaining files to HTML and defangs the emails by replacing hyperlinks with pop-up tooltips that display the URL on hover. We developed the Sanitise and Annotate Tool (SAT), a minimal browser-based tool for preparing PII-free and interactive email samples for use in *Phooled!*. It provides a GUI for replacing Personally Identifiable Information (PII) in an email's content with placeholders. The original recipient's name and email address are each replaced with a placeholder. Thus, a PII-free email sample is created while preserving the rest of the content and appearance of the email. When the email is shown during gameplay, these placeholders are populated with the player's own details, as described in Section 4.3.1, such that the simulated email is personalised for each player.

Using SAT, we manually classified each email, then identified and highlighted each phishing cue inline. We explicitly marked "spam" emails as Safe, since we only consider Phishing emails to be malicious. These annotations were used to provide interactive feedback in the game. To ensure the reliability of the ground truth labels, the classification and cue annotation of each email was carried out by a group of four researchers, with each sample independently labelled by two of them. When two annotators disagreed, the discrepancy was resolved through group discussion among the annotators to reach a consensus, which was taken as the ground truth. It should be noted that consensus was reached readily on classification, whereas identifying specific phishing cues within an email prompted more discussion. Finally, these labelled email samples were stored on the Phishducation platform and made accessible to the *Phooled!* game engine.

4.4.2 Justification for Selection of Cues

The annotation of attributes for these emails was based on a set of nine attributes found in the existing literature, as shown in Table 4.3. One of the key challenges of email labelling was identifying and marking cues in Safe emails, as current literature focuses

Table 4.3: Phishing cues used in *Phooled!*, as described in-game to players

Cue	Description	References
Link Mismatch	An unexpected or spoofed link with a URL that is malicious	[14, 27, 74]
Urgent Language	Urgent and pressuring language designed to provoke immediate action	[4, 27, 74, 87]
Image as Text	Use of text within images to evade text-based detection methods	[12, 31]
Spelling/Grammatical Errors	Presence of spelling and grammatical errors commonly found in phishing emails	[14, 27, 74]
Unaffiliated From Address	From address that is not affiliated with an organisation	[58, 59, 74]
Unlikely Reward	Promise of an unlikely reward	[4, 74]
Unprofessional Appearance/Tone	Email appearance or tone that is not expected from the sending party	[59, 74]
Sender Address Mismatch	From address that is unexpected or spoofed	[58, 74]
Subject Line Keywords	Suspicious keywords in the subject line typically used in phishing	[87]
General Phishing Language	General language patterns that may suggest phishing intent	Proposed
Money-Related Discussion	Discussion related to money or financial transactions	Proposed

solely on phishing cues present in Phishing emails. During group annotation consensus, two additional cues were identified that might be beneficial to teach during training. The first was a more cautious approach to "Unlikely Reward" called "Money-Related Discussion". It was designed to prime users to be alert at any mention of money in an email.

We also created a blanket category "General Phishing Language" for elements of an email that arouse suspicion or a general feeling of "trickiness," without any articulable reason. This categorisation was meant to act as a catch-all for this suspicion. The group also categorised each email into three levels of difficulty: Easy, Medium, and Hard, determined by group annotation consensus on the perceived difficulty of the email. These annotations were then embedded into the HTML file and then stored in the persistent database for

distribution during game sessions of *Phooled!*. A selection of these labelled samples, as presented to players, is shown in Appendix D.

Chapter 5

Phooled! Game Experimental Results and Discussion

Sixty-four participants took part in our study across three cohorts and treatment conditions. The 28 participants in the Collaborative treatment were batched into 10 groups of 2–5 players: 3 groups from HS Students, 3 from IT Professionals, and 4 from Remote participants. A breakdown of participants across cohorts and treatment conditions is presented in Table 5.1.

We begin by examining whether participants improved their phishing detection performance after playing *Phooled!* (Section 5.1) and, if so, whether individual characteristics predicted that improvement (Section 5.2). We then turn to the game itself, assessing the mechanics of *Phooled!* (Section 5.3) and determining in-game phishing detection difficulty (Section 5.4). We conclude with an analysis of participant feedback (Section 5.5).

5.1 Phishing Detection Improvement in Learners

Using the three performance measures described in Section 4.2 (Score, Duration, and IES), we analysed the effect of playing *Phooled!* on phishing detection. We first analysed the

Table 5.1: Breakdown of participants across cohort and treatment conditions

Date	**Cohort**	**Mode**	**Treatment**	***n***
Aug. 2024	IT Professionals	In-Person	Single-Player	5
			Collaborative	9
			Subtotal	*14*
Nov. 2024	High School Students	In-Person	Single-Player	7
			Collaborative	8
			Subtotal	*15*
Jul. 2025	Prolific/Online	Remote	Control	11
			Single-Player	12
			Collaborative	12
			Subtotal	*35*
			Total	**64**

participants' performance on the Pre-Test to establish their baseline phishing detection ability prior to any training. Table 5.2 summarises the overall baseline performance. Participants scored a mean test score of 7.47 ($SD = 1.65$), with a mean duration of 360 seconds ($SD = 177$). The resulting mean IES, which takes into account both speed and accuracy, was 50.03 ($SD = 26.23$).

Importantly, the distributions of all three metrics were right-skewed, with means higher than medians. This suggests that a subset of participants were at an extreme: some participants were much more accurate at detecting phishing than others, while others took considerably longer time or were less efficient than the average. This indicates that the average participant, prior to any training, already exhibited moderate phishing detection ability, which is consistent with that reported in prior studies such as [44, 90].

The skewed nature of the distributions was confirmed with normality tests, specifically

Table 5.2: Pre-Test descriptive statistics ($n = 64$)

Measure	**Mean**	***SD***	**Median**
Score (out of 10)	7.47	1.65	7.00
Duration (s)	360.16	176.80	354.50
IES	50.03	26.23	46.14

the Shapiro-Wilk test, which found that all three measures deviated significantly from normality: Score ($W = 0.95$, $p = .009$), Duration ($W = 0.92$, $p < .001$), and IES ($W = 0.89$, $p < .001$). We therefore adopted an Analysis of Covariance (ANCOVA) approach [23] using mixed models, controlling for pre-training performance and individual variations when comparing treatment effects. The ANCOVA model for each performance measure is derived from the following generalised form:

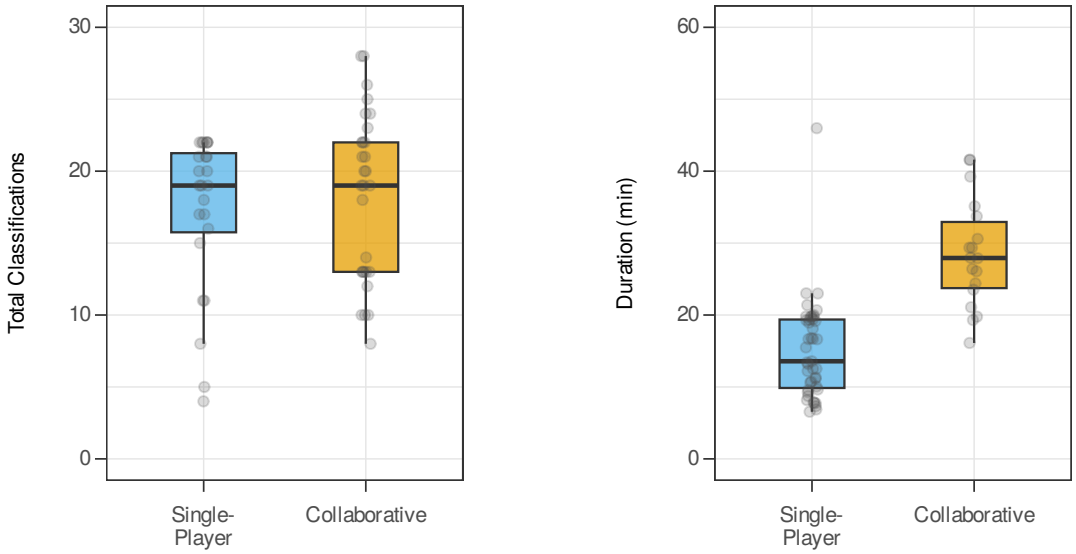
$$Y_{\text{post}_t} \sim Y_{\text{pre}} + \text{Treatment} + \text{Cohort} + t \tag{5.1}$$

where:

- Y is the performance measure, $Y \in \{\text{Score, Duration, IES}\}$
- Y_{pre} is the Pre-Test performance measure, used as the baseline
- t indexes the Post-Test assessment, $t \in \{1, 2\}$

The test score values are bounded between 0–10, compatible with a Linear Mixed Model (LMM). In contrast, the test duration and IES values are strictly positive, justifying the use of a Gamma Generalised Linear Mixed Model (GLMM) with a log link. The Duration and IES models included random intercepts and slopes for each participant across tests to account for individual variances in improvement during the study. The Score model included random intercepts only, as there was insufficient variance in the data to justify random slopes.

First, we verified that participants in both non-Control treatment conditions received equivalent amounts of learning material as intervention. As shown in Figure 5.1(a), participants interacted with (i.e., classified) a similar number of email samples in both Single-Player ($n = 24$, $M = 17.2$, $SD = 5.51$) and Collaborative ($n = 29$, $M = 18.3$, $SD = 5.84$) games. Q–Q plots and Shapiro-Wilk tests indicated significant deviations from normality, as shown in Appendix C.1. Therefore, a Wilcoxon rank-sum test confirmed no significant difference in the number of email samples classified ($W = 314$, $p = .50$). Although both treatment conditions involved similar exposure to learning material, they differed in game length.



(a) (b)

Figure 5.1: Boxplots of total classifications and game duration by treatment condition

As shown in Figure 5.1(b), Single-Player games were shorter ($M = 15.1$ min, $SD = 6.96$) than Collaborative games ($M = 28.5$ min, $SD = 7.50$). Q–Q plots and Shapiro-Wilk tests indicated significant deviations from normality, as shown in Appendix C.2. A Wilcoxon rank-sum test confirmed the difference was significant ($W = 59$, $p < .001$, $r = 0.66$), as we expected, given the additional time required for game setup, group coordination, and discussion in Collaborative games.

This establishes that both treatment conditions were exposed to similar amounts of learning material, with only game durations differing due to group gameplay. The remainder of this section tests H1 and H2 by analysing participant phishing detection improvement after training with *Phooled!* across these three performance measures: accuracy (Section 5.1.1), speed (Section 5.1.2), and efficiency (Section 5.1.3).

5.1.1 Effect on Phishing Detection Accuracy

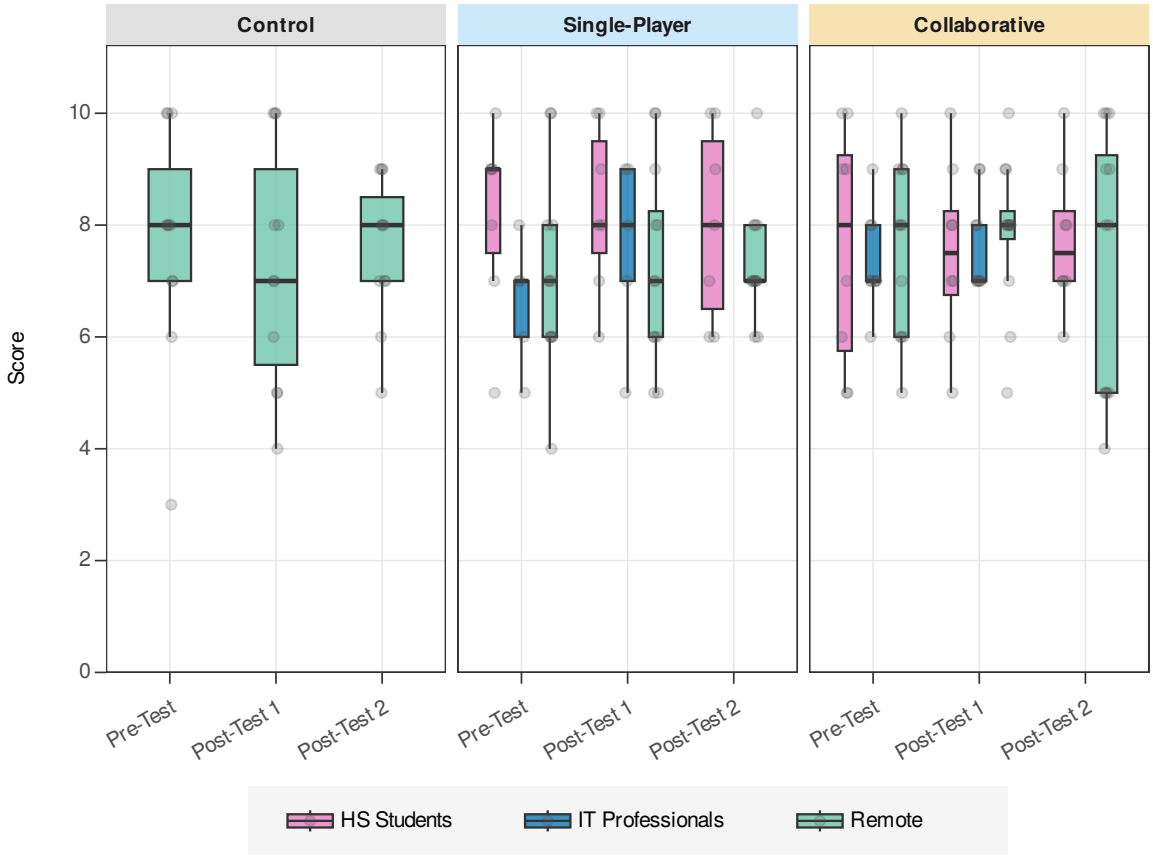


Figure 5.2: Boxplot of Test Score values

Figure 5.2 shows the distribution of test scores by cohort and treatment condition across all three tests. The mean accuracy was 7.47 (SD = 1.65) for the Pre-Test, 7.59

($SD = 1.57$) for Post-Test 1, and 7.54 ($SD = 1.58$) for Post-Test 2. Accordingly, overall median accuracy remained approximately constant across all tests. Moreover, as shown in Figure 5.2, no visible separation is observed between any treatment condition or cohort. Q–Q plots and Shapiro-Wilk tests by cohort and treatment condition are provided in Appendix C.13.

We fitted the ANCOVA model from Equation 5.1 with Score as the dependent variable, and with DHARMa [36] residual diagnostics confirming adequate model fit. The model explained 39.4% of total variance in the data (conditional $R^2 = .394$), of which 10.5% was due to fixed effects alone (marginal $R^2 = .105$), indicating that individual differences between participants accounted for most of the variance in the data. After controlling for the baseline Pre-Test score, which was a significant covariate ($\chi^2(1) = 8.74$, $p = .003$), treatment had no significant effect on accuracy ($\chi^2(2) = 0.13$, $p = .936$, $\eta_p^2 < .01$).

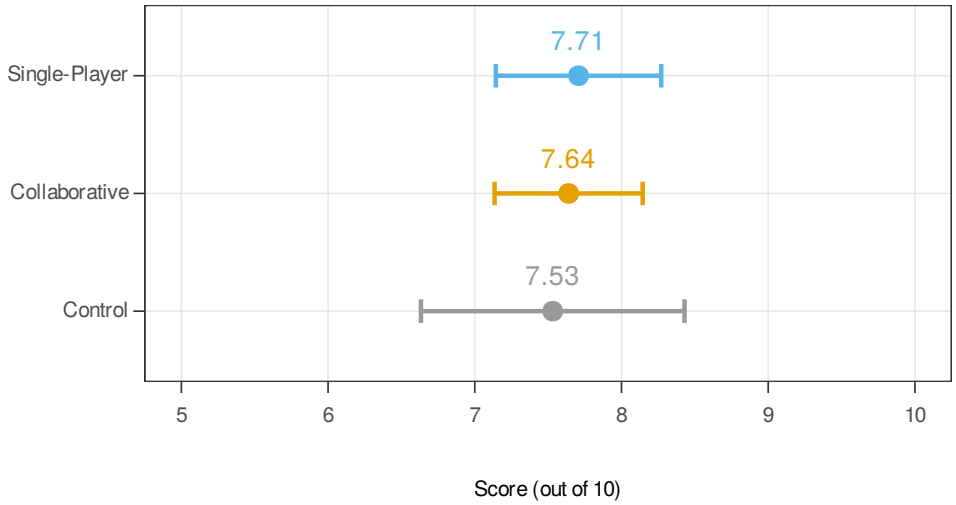


Figure 5.3: EMMs for post-training test scores by treatment condition (error bars indicate 95% CIs)

As shown in Figure 5.3, Estimated Marginal Means (EMMs) for post-training accuracy were similar across treatment conditions: Control ($M = 7.53$, $SE = 0.45$), Single-Player ($M = 7.71$, $SE = 0.28$), and Collaborative ($M = 7.64$, $SE = 0.25$). Therefore, after controlling for pre-training performance, no significant differences in accuracy were observed

across treatment conditions. Thus, H1 and H2 are not supported for this measure.

5.1.2 Effect on Phishing Detection Speed

Figure 5.4 shows the distribution of test duration by cohort and treatment condition across all three tests. The mean duration decreased steadily from 360 seconds ($SD = 177$) for the Pre-Test, to 275 seconds ($SD = 176$) for Post-Test 1, and 207 seconds ($SD = 122$) for Post-Test 2, indicating that participants completed the assessments faster over time. The high SD values reinforce the right-skewed nature of the distribution, as some participants took much longer to complete the assessments. The boxplot suggests a visual trend of declining test durations for the Single-Player and Collaborative conditions compared to the Control group. Q–Q plots and Shapiro-Wilk tests by cohort and treatment condition are provided in Appendix C.14.

We fit the ANCOVA model from Equation 5.1 with Duration as the dependent variable, and with DHARMa residual diagnostics confirming adequate model fit. The model explained 66.8% of total variance in the data (conditional $R^2 = .668$), of which 46.8% was due to fixed effects alone (marginal $R^2 = .468$). After controlling for the baseline Pre-Test duration, which was a significant covariate ($\chi^2(1) = 26.46$, $p < .001$), treatment had a significant effect on duration ($\chi^2(2) = 9.01$, $p = .011$).

As shown in Figure 5.5, EMMs for post-training duration indicated that Collaborative participants completed tests the fastest ($M = 181$ s, $SE = 18.8$), followed by Single-Player ($M = 228$ s, $SE = 23.7$), and finally Control ($M = 254$ s, $SE = 38.7$). The follow-up one-tailed pairwise comparisons showed that Collaborative participants completed their Post-Test assessments 29% faster than Control participants ($z = -2.60$, $p = .005$) and 21% faster than Single-Player participants ($z = -2.36$, $p = .009$), supporting H2. This decrease in test duration did not further accelerate between Post-Test 1 and Post-Test 2 ($\chi^2(2) = 1.16$, $p = .559$), indicating that the advantage of Collaborative training was established by the first game session and that the second game session did not widen the gap between the treatments.

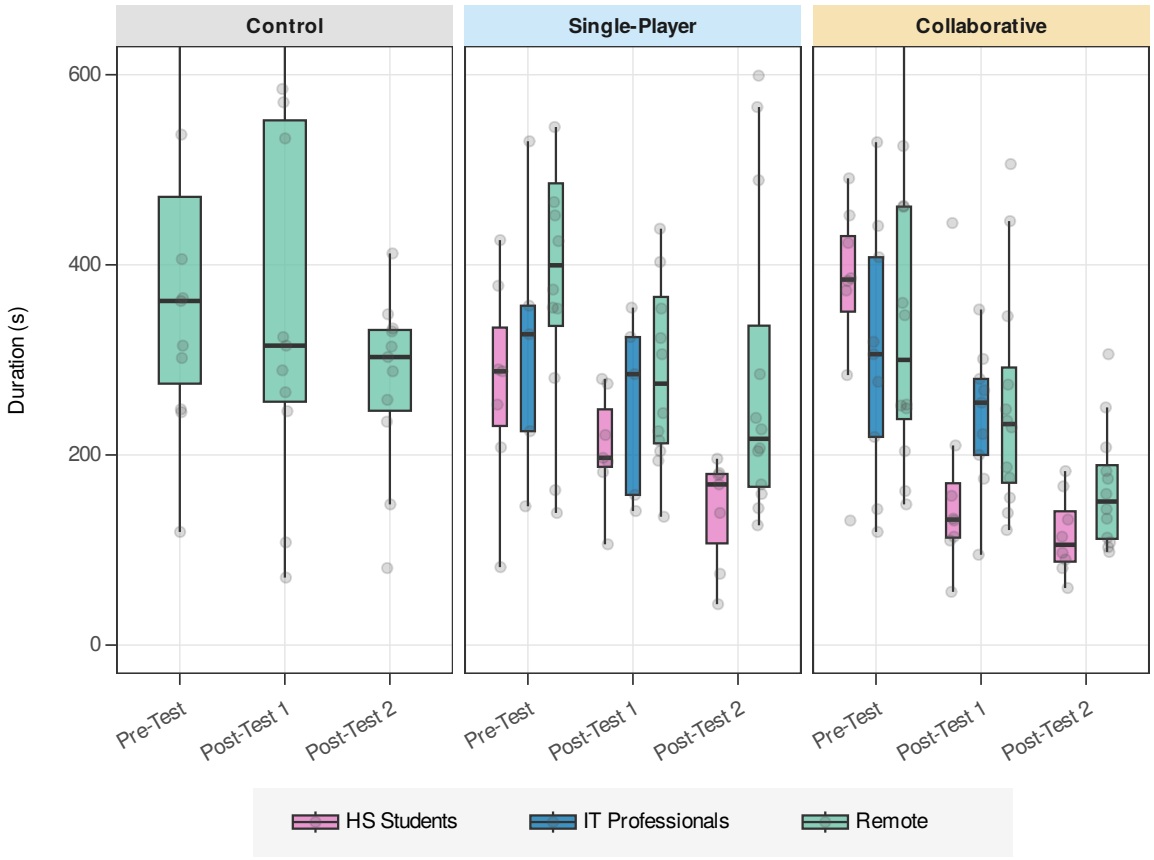


Figure 5.4: Boxplot of Test Duration values

In contrast, Single-Player participants were only 10% faster than Control ($z = -0.81$, $p = .208$), suggesting no significant reduction in time compared to receiving no treatment. This is due to the fact that Control participants, who received no treatment, still reduced their median test duration from 362 seconds in the Pre-Test to 303 seconds in Post-Test 2, a 16% reduction. This suggests an underlying practice effect from repeated testing. Single-Player participants, unlike Collaborative participants, did not improve sufficiently beyond this practice effect for us to detect a significant difference compared to Control, and thus H1 is only partially supported for this measure.

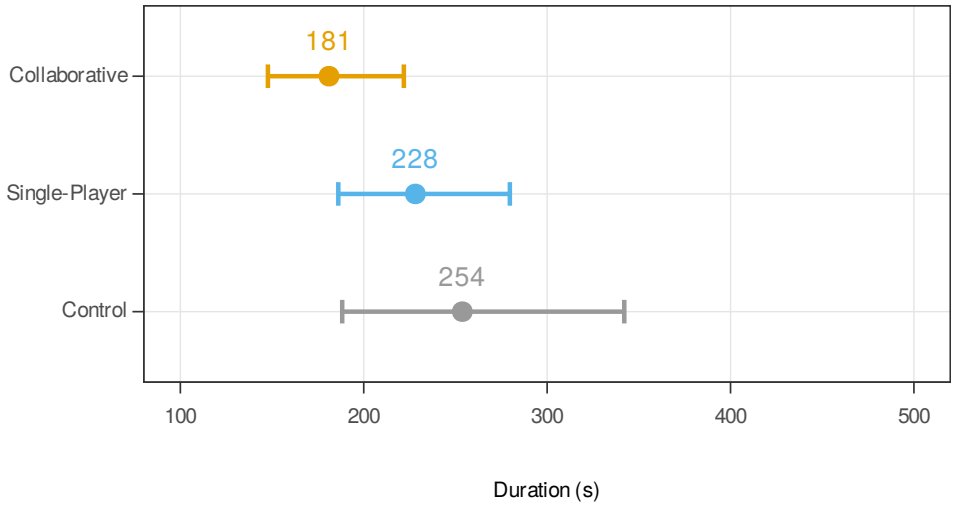


Figure 5.5: EMMs for post-training test duration by treatment condition (error bars indicate 95% CIs)

5.1.3 Effect on Phishing Detection Efficiency

As described in Section 4.2, the IES measure helps us by analysing the speed-accuracy tradeoff, ensuring that improvements in speed did not come at the cost of decreasing accuracy. Figure 5.6 shows the distribution of IES (where lower values indicate better efficiency) by cohort and treatment condition across all three tests. The mean IES decreased steadily from 50.03 ($SD = 26.23$) for the Pre-Test, to 37.76 ($SD = 25.03$) for Post-Test 1, and 28.01 ($SD = 15.81$) for Post-Test 2, indicating that participants consistently improved their phishing detection efficiency across the tests. The boxplot shows a visual trend of improving efficiency across all three conditions, with the Collaborative and Single-Player conditions showing a steeper improvement than the Control group. Q–Q plots and Shapiro-Wilk tests by cohort and treatment condition are provided in Appendix C.15.

We fit the ANCOVA model from Equation 5.1 with IES as the dependent variable, and with DHARMa residual diagnostics confirming adequate model fit. The model explained 61.6% of total variance in the data (conditional $R^2 = .616$), of which 42.1% was due to fixed effects alone (marginal $R^2 = .421$). After controlling for the baseline Pre-Test IES, which

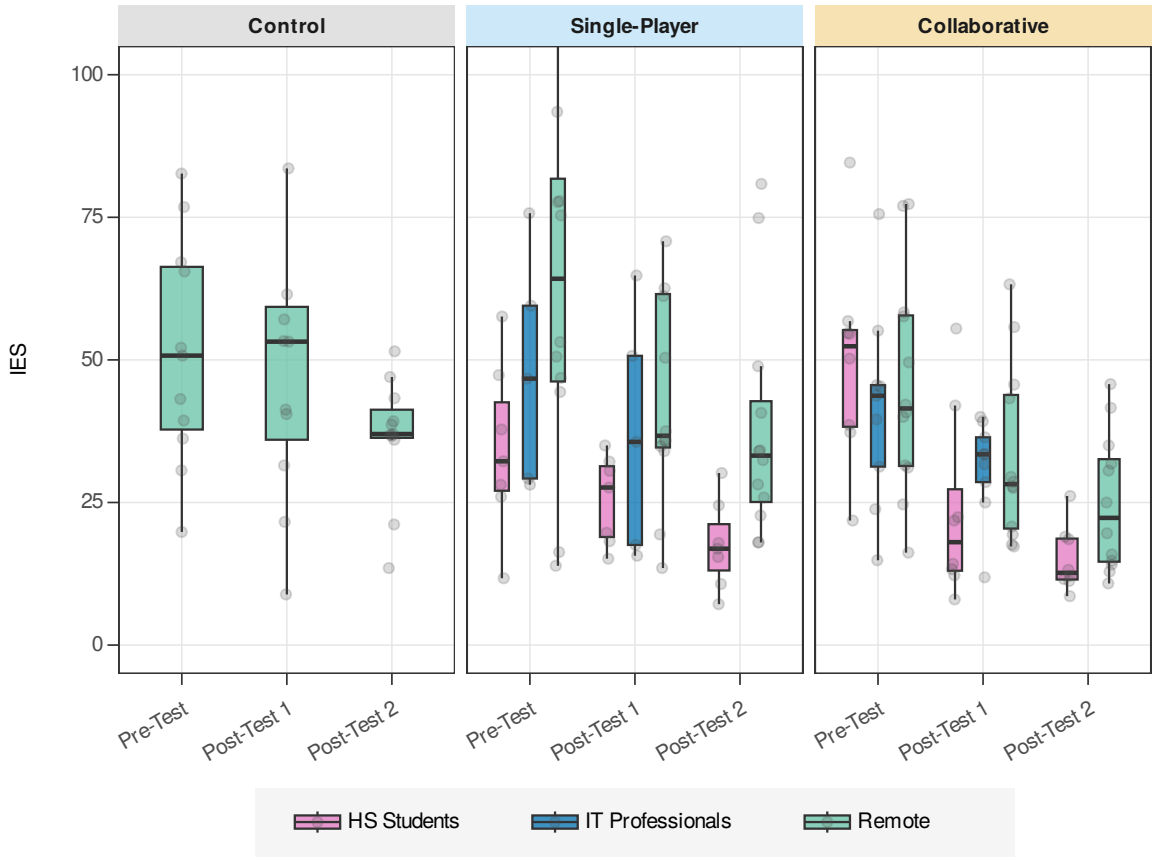


Figure 5.6: Boxplot of Test Inverse Efficiency scores (IES), where lower values indicate better efficiency.

was a significant covariate ($\chi^2(1) = 9.58$, $p = .002$), treatment had a significant effect on IES ($\chi^2(2) = 8.50$, $p = .014$). As shown in Figure 5.7, EMMs for post-training efficiency indicated that Collaborative participants were the most efficient ($M = 24.1$, $SE = 1.83$), ahead of both Single-Player ($M = 30.2$, $SE = 2.61$) and Control ($M = 35.3$, $SE = 4.77$). The follow-up one-tailed pairwise comparisons showed that Collaborative participants were 32% more efficient than Control participants ($z = -2.65$, $p = .004$) and 20% more efficient than Single-Player participants ($z = -2.06$, $p = .020$), supporting H2.

In comparison, Single-Player participants were only 15% more efficient than Control

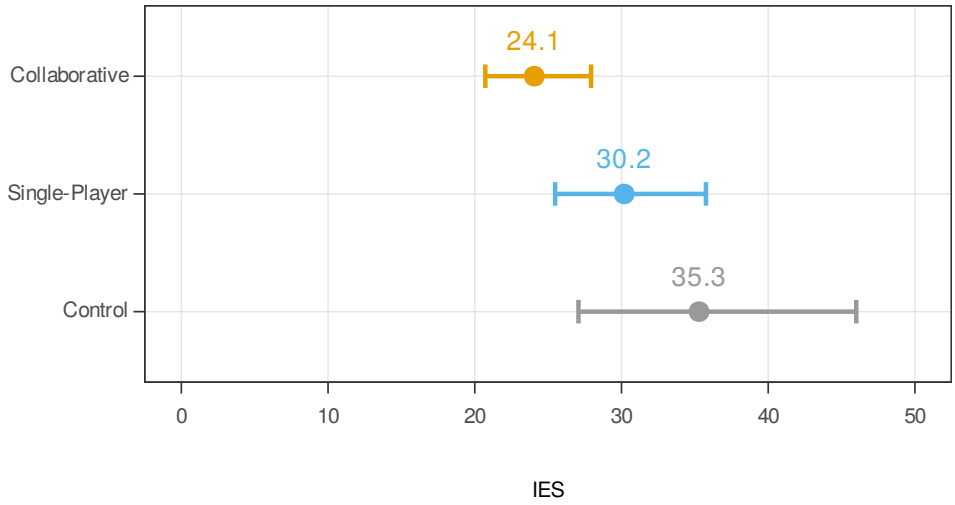


Figure 5.7: EMMs for post-training IES by treatment condition (error bars indicate 95% CIs)

($z = -1.05$, $p = .147$), suggesting no significant improvement in efficiency compared to receiving no treatment. As with Duration, this advantage was established after the first game session itself ($\chi^2(2) = 0.24$, $p = .888$). Control participants also improved their IES from a median of 50.8 in the Pre-Test to 37.0 in Post-Test 2, a 27% reduction, consistent with the practice effect reported in Section 5.1.2. Single-Player participants did not improve beyond this practice effect, and thus H1 is only partially supported for this measure.

5.1.4 Summary of Phishing Detection Improvement

We observed that while accuracy remained relatively the same after training, both speed and efficiency improved with Collaborative play using *Phooled!*. Participants who played Collaborative *Phooled!* improved their overall phishing detection speed by 29% and their efficiency by 32% compared to the Control group. Since IES is a measure of the speed-accuracy tradeoff, we could thus confirm that improvements in phishing detection speed through Collaborative play did not come at the cost of a decrease in detection accuracy. This lack of accuracy improvement, despite the gains in speed and efficiency, is likely due

to the fact that participants already began with high detection accuracy (a mean Pre-Test score of around 7.5 out of 10), leaving little room for further improvement. Their baseline detection times, by contrast, were considerably longer (a mean of 360 seconds), leaving much more room to improve.

We did not find any significant evidence to show that Single-Player *Phooled!* gameplay differed from the Control group for any measure. H1 is therefore partially supported: training with *Phooled!* improved speed and efficiency but not accuracy, and only for Collaborative training. As a corollary, H2 is thus supported, with Collaborative training outperforming Single-Player on both speed and efficiency.

Table 5.3: Average False Negative Rate (FNR) and False Positive Rate (FPR) for each assessment

	FNR		**FPR**	
Test	**Mean**	***SD***	**Mean**	***SD***
Pre-Test	0.17	0.18	0.34	0.27
Post-Test 1	0.14	0.17	0.35	0.26
Post-Test 2	0.17	0.18	0.33	0.26

To investigate the non-improvement in phishing detection accuracy further, we decomposed it into two constituents: the FNR and the FPR, as shown in Table 5.3. The overall FNR remained approximately the same throughout the study, with a mean of 0.16, indicating that participants were generally good at detecting phishing emails prior to training. On the other hand, the overall FPR was consistently higher, with a mean of 0.34, suggesting that participants' accuracy is dragged down by a frequent misclassification of Safe emails as Phishing. This asymmetry between Phishing and Safe email classification is explored further in Section 5.3.1 and Section 5.4, where we examine how participants processed Phishing and Safe emails differently during gameplay.

5.2 Predictors of Phishing Detection Improvement

Prior to treatment, participants completed a calibration survey that measured their motivation and social influence factors, based on the scales introduced in Chapter 2 and detailed in Section 4.2. To answer H3 and H4, we examine whether scales like MPA (Section 5.2.1), Test Anxiety (Section 5.2.2), Subjective Norms (Section 5.2.3), and instruction mode (Section 5.2.4) predicted phishing detection improvement.

Table 5.4: Descriptive statistics for calibration scales by cohort

Scale	HS Students ($n = 14$)	IT Professionals ($n = 14$)	Remote ($n = 35$)	Overall ($N = 63$)
MPA	5.26 (0.72)	5.60 (0.87)	6.03 (0.69)	5.76 (0.80)
Test anxiety	3.67 (1.18)	2.86 (1.46)	4.14 (1.81)	3.75 (1.67)
SN1	3.90 (0.76)	4.02 (1.33)	3.79 (0.93)	3.87 (0.99)
SN2	5.14 (1.08)	5.96 (1.84)	5.11 (1.71)	5.31 (1.64)
Heuristic	3.00 (0.85)	3.05 (1.06)	2.95 (0.85)	2.98 (0.88)
Systematic	3.76 (0.90)	4.33 (0.64)	4.05 (1.12)	4.05 (0.99)

Note. Scales were measured on 1–5 or 1–7 Likert scales (see Table 4.2), given above as M (SD).

Table 5.4 summarises the mean and standard deviation values of each calibration scale by cohort. The IT Professionals cohort showed the most notable differences from the other cohorts, reporting the lowest Test Anxiety ($M = 2.86$) compared to HS Students ($M = 3.67$) and Remote ($M = 4.14$). The SN2 scale values were also highest among IT Professionals ($M = 5.96$), compared to HS Students ($M = 5.14$) and Remote ($M = 5.11$). The Remote cohort showed notably higher MPA ($M = 6.03$) than HS Students ($M = 5.26$) and IT Professionals ($M = 5.60$). These cohort-level differences justified the inclusion of cohort interaction terms in the analysis of the survey data.

The Heuristic ($M = 2.98$) and Systematic ($M = 4.05$) scales were relatively uniform across cohorts. Among our calibration scales, Test Anxiety ($SD = 1.67$) and SN2 ($SD = 1.64$) showed the highest individual variability, while the MPA ($SD = 0.80$) and Heuristic

($SD = 0.88$) scales had the lowest spread of values. The normality diagnostics for all calibration scales are provided in Appendix C.

$$\text{IES}_{\text{post}_t} \sim \text{IES}_{\text{pre}} + \text{Treatment} + \text{Cohort} + t + (\text{Treatment} + \text{Cohort}) \times \mathbf{S} \tag{5.2}$$

To examine whether participants' scores on these scales predicted post-training IES, we initially extended the Gamma GLMM from Equation 5.1 by adding our six calibration scales, namely MPA, Test Anxiety, SN1, SN2, Heuristic, and Systematic ($N = 63$). These scales are represented in Equation 5.2 above by the vector $\mathbf{S}$. One participant did not complete the calibration survey and was subsequently excluded from this analysis. The per-participant variability for the GLMM was near zero, therefore, the random intercepts were excluded to form a Generalised Linear Model (GLM) instead.

We also included interaction terms, represented in brackets in the equation, to account for treatment and cohort-level scale differences. Based on the results from Section 5.1, we proceeded with IES as the dependent variable for this and subsequent sections in our analysis, as it best captures overall phishing detection performance. By using pre-training IES as a covariate, we could isolate the effect of our scales to predict how they affect improvement beyond pre-training performance.

Table 5.5 above presents the results from the model fitted on Equation 5.2. Lower b values correspond to lower post-training IES, indicating better performance. The model explained 67.9% of the variance in the post-training IES data ($R^2 = .679$), with DHARMa residual diagnostics confirming adequate model fit. The Heuristic and Systematic scales showed no significant effects or interactions for any condition and are excluded from analysis in this section. The effect of each remaining scale on IES improvement after training is examined in the following subsections.

Table 5.5: Fixed effects from the Gamma GLM predicting post-training IES ($N = 63$).

Scale	b	SE	95% CI		z	p
MPA	−0.63	0.25	[−1.13,	−0.13]	−2.49	.013*
Test Anxiety	−0.19	0.13	[−0.45,	0.06]	−1.47	.140
SN1	0.86	0.34	[0.20,	1.53]	2.56	.011*
SN2	−0.67	0.17	[−0.99,	−0.34]	−4.03	< .001***
Heuristic	0.37	0.25	[−0.12,	0.85]	1.47	.142
Systematic	0.45	0.27	[−0.07,	0.97]	1.68	.093

*$p < .05$. **$p < .01$. ***$p < .001$.

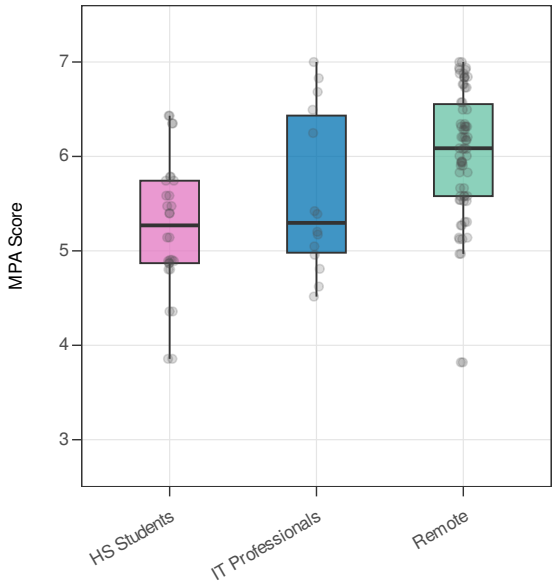


Figure 5.8: Boxplot of MPA scores by cohort

5.2.1 Effect of Participant Motivation

The distribution of MPA scores by cohort is shown in Figure 5.8, with Q–Q normality plots in Appendix C.3. Higher MPA scale values predicted lower post-training IES ($b = -0.63$,

$p = .013$). Interpreted on the scale's 1–7 range, this means that each unit increase in MPA was associated with approximately 47% lower post-training IES values, with a CI of 12% to 68%.

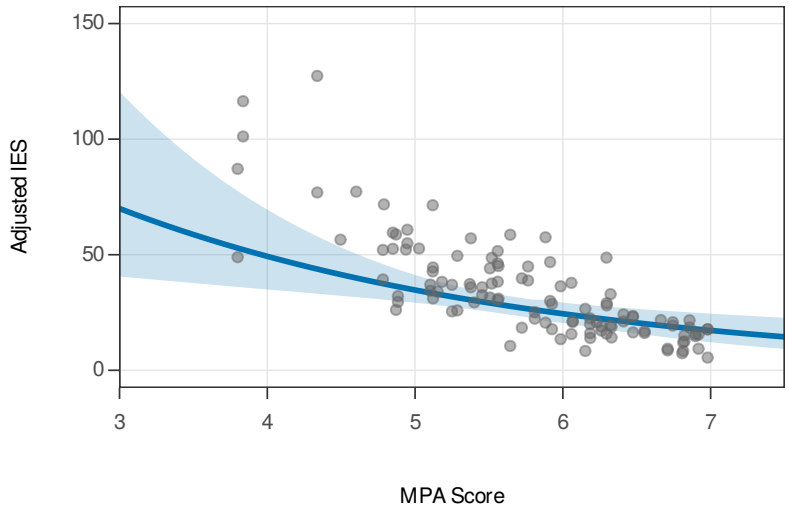


Figure 5.9: Scatterplot of post-training IES by MPA score, adjusted for covariates (95% CI band)

This relationship between MPA values and post-training IES is visualised in Figure 5.9. The figure shows a clear downward trend, with Adjusted IES decreasing as MPA scores increase, consistent with the predicted negative effect. In this figure, Adjusted IES refers to IES values adjusted for the influence of covariates, including pre-training IES. Neither the cohort ($\chi^2(2) = 5.08$, $p = .079$) nor the treatment ($\chi^2(2) = 4.94$, $p = .084$) interactions were significant, suggesting that the effect of MPA did not differ among cohort and treatment conditions. Therefore, H4 is supported: higher Motivation and Perceived Ability predicted greater improvement in phishing detection efficiency after playing *Phooled!*, regardless of whether Single-Player or Collaborative mode was played.

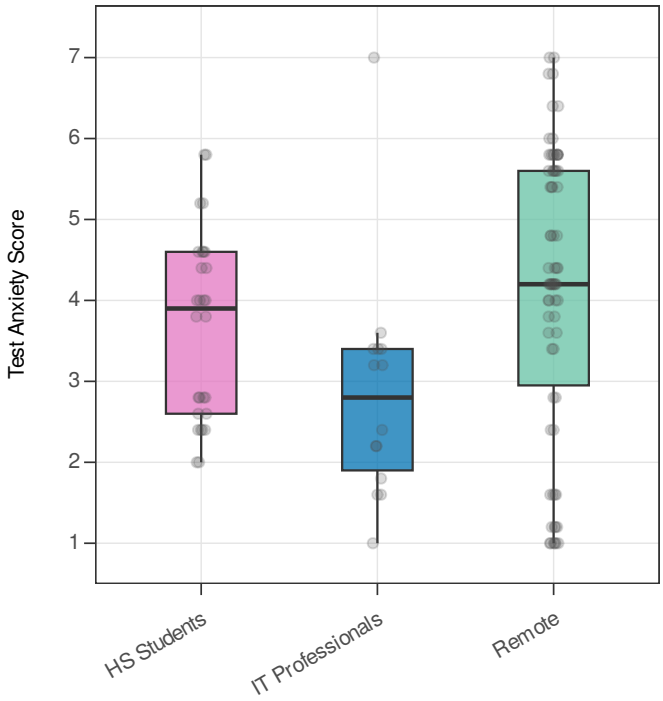
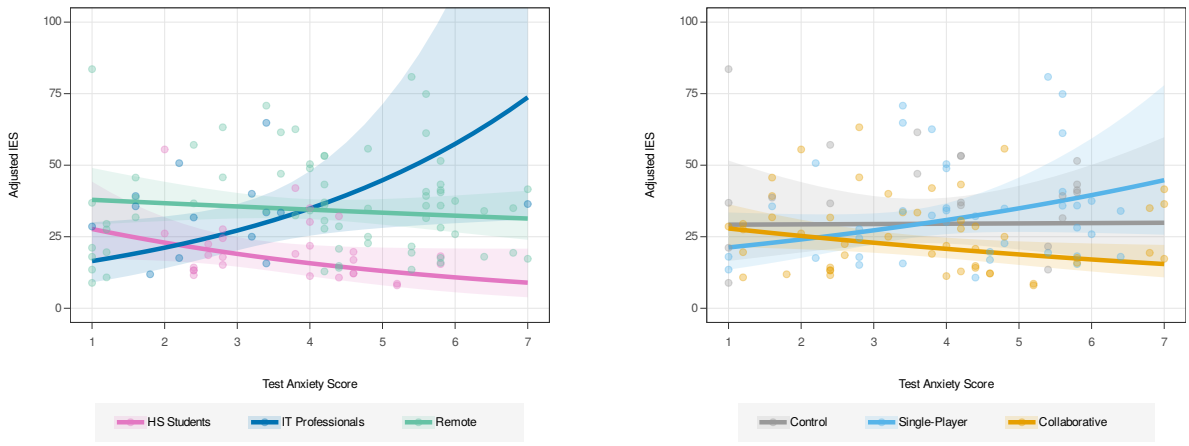


Figure 5.10: Boxplot of Test Anxiety scores by cohort

5.2.2 Effect of Test Anxiety

The distribution of Test Anxiety scores by cohort is shown in Figure 5.10, with Q–Q normality plots in Appendix C.4. The fixed effect of Test Anxiety on IES improvement was not statistically significant ($b = -0.19$, $p = .140$). This is due to the fact that the overall effect of Test Anxiety was masked by opposing effects across cohort ($\chi^2(2) = 10.36$, $p = .006$) and treatment ($\chi^2(2) = 6.72$, $p = .035$) interactions, as shown in Figure 5.11.

In Figure 5.11(a), IT Professionals have a significantly steeper positive slope than the other cohorts. This indicates that higher Test Anxiety scores were associated with worse efficiency for this cohort, to a greater degree than for HS Students and Remote participants. In contrast, Figure 5.11(b) shows that the Collaborative condition exhibits a significant negative slope. This suggests that higher Test Anxiety scores were associated with better efficiency for Collaborative participants, an effect not observed in Control and Single-Player



(a) (b)

Figure 5.11: Adjusted post-training IES by Test Anxiety score, split by cohort (a) and treatment conditions (b) (95% CI bands)

participants.

Specifically, for IT Professionals, higher Test Anxiety scores predicted worse IES ($b = 0.25$, CI $[0.03, 0.47]$), an effect that differed significantly from HS Students ($z = 3.21$, $p = .004$), who showed a marginal but non-significant trend in the opposite direction ($b = -0.19$, CI $[-0.39, 0.01]$). This aligns with our findings in Section 5.4.3, suggesting that test anxiety may interfere with leveraging one's existing knowledge to detect phishing under duress. For Remote participants, there was no effect observed in either direction ($b = -0.03$, CI $[-0.11, 0.05]$). Among treatment conditions, Collaborative gameplay improved IES for those with higher Test Anxiety scores ($b = -0.10$, CI $[-0.19, -0.01]$), differing significantly from the Single-Player condition ($z = 2.59$, $p = .026$). The other treatment conditions did not show any significant effect.

5.2.3 Effect of Subjective Norms

The boxplot of subjective norms scales by cohort is shown in Figure 5.12, with Q–Q normality plots in Appendix C.12. Interestingly, the two Subjective Norms (SN) scales

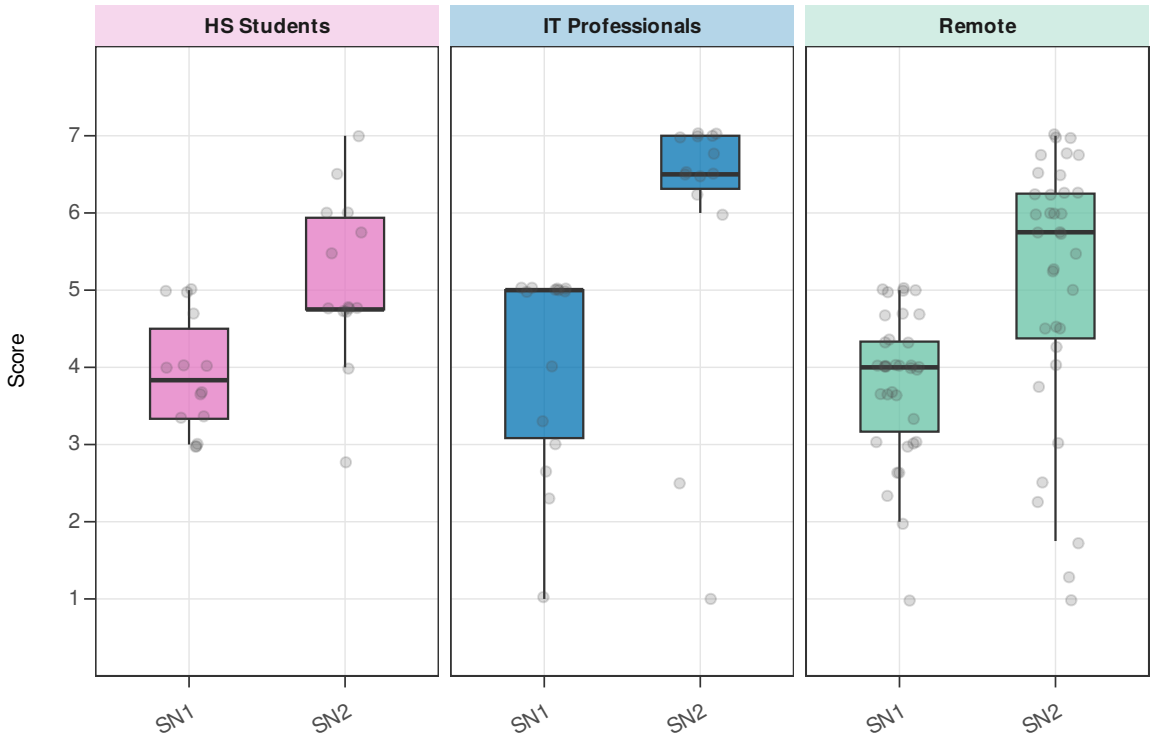


Figure 5.12: Boxplot of Subjective Norms scale values by cohort

showed opposing fixed effects. Higher SN1 values were associated with worse IES scores ($b = 0.86$, $p = .011$), while the opposite was true for higher SN2 values ($b = -0.67$, $p < .001$).

These opposing effects may reflect the differences in the subjective norms scales. SN1 measures general peer influence [42], while SN2 measures influence due to authority [71]. These effects were further influenced by the treatment condition, as shown in Figure 5.13, with the fixed effects for both SN1 ($\chi^2(2) = 13.84$, $p < .001$) and SN2 ($\chi^2(2) = 19.42$, $p < .001$) being dependent on significant interactions.

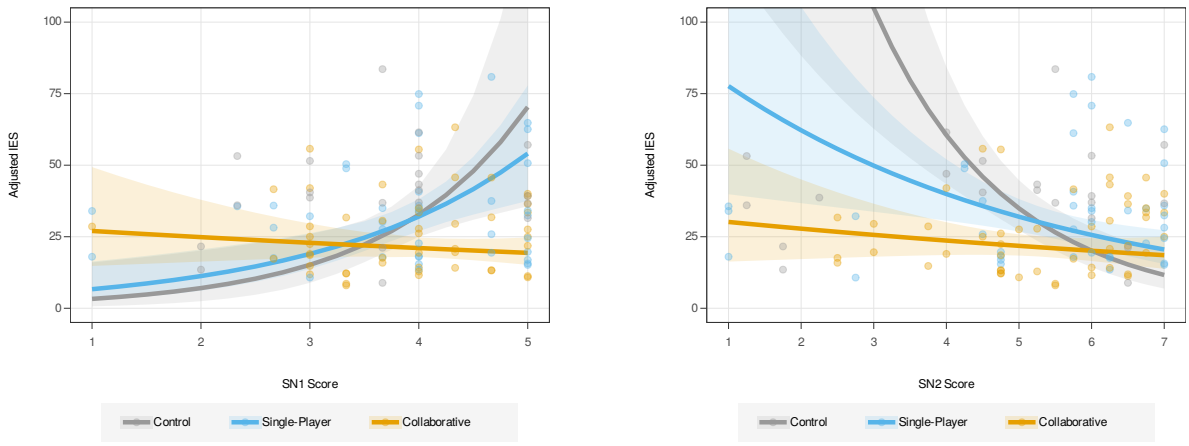
Specifically, for SN1, we observed a harmful effect on IES in Control ($b = 0.77$, CI $[0.22, 1.32]$) and Single-Player ($b = 0.52$, CI $[0.22, 0.83]$) treatments. This harmful effect was not present in Collaborative training ($b = -0.08$, CI $[-0.27, 0.11]$). This difference

was significant between Collaborative training and both the Control group ($z = 2.83$, $p = .013$) and Single-Player ($z = 3.17$, $p = .004$) training. This suggests that peer influence had a harmful effect on phishing detection efficiency improvement, particularly when collaborative learning is not involved.

In contrast, we observed that for SN2, there was a beneficial effect on IES in the Control group ($b = -0.55$, CI $[-0.78, -0.32]$) and Single-Player ($b = -0.22$, CI $[-0.37, -0.08]$) training. This effect was again absent in the Collaborative treatment ($b = -0.08$, CI $[-0.22, 0.06]$). This difference was significant between Collaborative training and the Control group ($z = 4.35$, $p < .001$) but not the Single-Player condition ($z = 1.51$, $p = .285$). This suggests that users who valued organisational authority showed greater improvement in phishing detection efficiency, particularly when collaborative learning is not involved. In essence, Collaborative gameplay neutralised the effect of subjective norms on post-training IES.

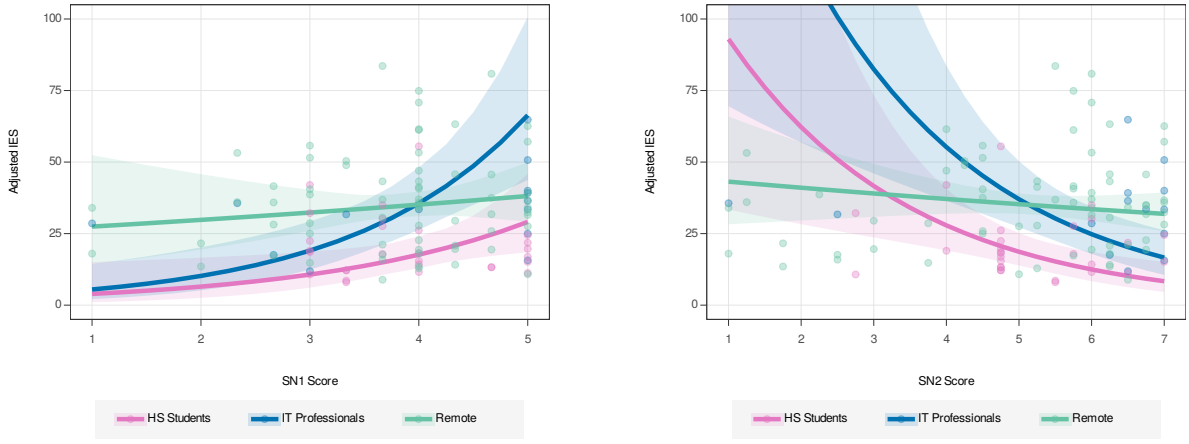
Similarly, the cohort interactions for SN1 ($\chi^2(2) = 9.54$, $p = .008$) and SN2 ($\chi^2(2) = 10.20$, $p = .006$) were also significant, as shown in Figure 5.14. The harmful effect of SN1 was present in IT Professionals ($b = 0.62$, CI $[0.32, 0.93]$) and HS Students ($b = 0.50$, CI $[0.09, 0.92]$) only and not in Remote participants ($b = 0.08$, CI $[-0.14, 0.30]$). The difference between IT Professionals and Remote was statistically significant ($z = 3.07$, $p = .006$).

A similar pattern was observed for the beneficial effect of SN2: IT Professionals ($b = -0.40$, CI $[-0.61, -0.19]$) and HS Students ($b = -0.40$, CI $[-0.65, -0.15]$) benefited from higher SN2 values while Remote participants ($b = -0.05$, CI $[-0.15, 0.05]$) did not. For SN2, both the IT Professionals ($z = 2.81$, $p = .014$) and HS Students ($z = 2.54$, $p = .030$) cohorts were significantly different from Remote, confirming that both scales were dependent on instruction mode, in-person or Remote. Therefore, H3 is supported: subjective norms predicted post-training IES, however, both scales had varying effects dependent on treatment condition and instruction mode.



(a) (b)

Figure 5.13: Adjusted post-training IES by SN1 (a) and SN2 (b) scores, split by treatment condition (95% CI bands)



(a) (b)

Figure 5.14: Adjusted post-training IES by SN1 (a) and SN2 (b) scores, split by cohort (95% CI bands)

5.2.4 Effect of Instruction Mode

Although *Phooled!* was played online with no physical interaction between players, the HS Students and IT Professionals cohorts were part of an in-person study, while the Remote

cohort participated fully online. In Sections 5.2.2 and 5.2.3, we observed that scale interactions were dependent on instruction mode. In the same vein, we examined here whether instruction mode directly predicted post-training IES.

The effect of cohort on post-training IES was significant ($\chi^2(2) = 13.98$, $p < .001$). As shown in Figure 5.15, HS Students showed the most improvement after training ($M = 22.4$, $SE = 2.42$), followed by IT Professionals ($M = 32.5$, $SE = 4.42$), and finally Remote participants ($M = 35.3$, $SE = 2.33$). The significant fixed effect was driven solely by HS Students, who improved significantly more than IT Professionals ($p = .048$) and the Remote cohort ($p = .001$), while the IT Professionals and Remote cohorts did not differ ($p = .848$). Therefore, we cannot conclude that in-person instruction mode alone explains these differences, as they were specific to the HS Students cohort.

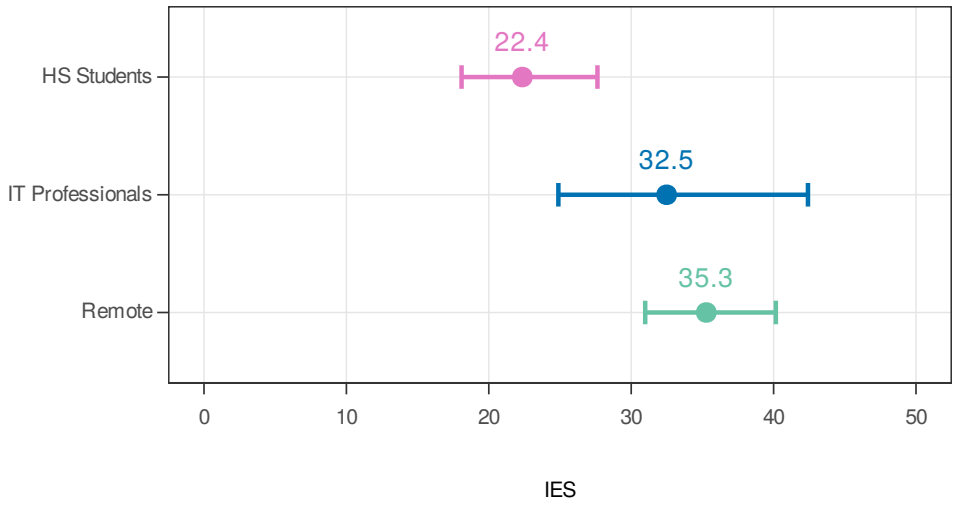


Figure 5.15: EMMs for post-training IES by cohort (error bars indicate 95% CIs)

5.2.5 Summary of Predictors of Phishing Detection Improvement

Of the six calibration scales examined, three of them, namely MPA, SN1, and SN2, directly predicted improvements in IES after training, while the Test Anxiety scale depended on

cohort and treatment interactions. Participants who scored higher on the Motivation and Perceived Ability scale went on to significantly improve their phishing detection efficiency regardless of cohort or treatment condition, supporting H4. The subjective norms scales had different effects: SN1 was associated with worse performance and SN2 with better performance, supporting H3.

However, these effects were not present in Collaborative training, which improved performance regardless of subjective norms. The effect of Test Anxiety varied by cohort and treatment condition, as participants who scored higher performed slightly better after Collaborative treatment. Finally, since cohort differences in IES improvement were driven solely by HS Students, we did not see any differences that could be attributed to instruction mode alone. Having identified which individual characteristics of participants helped improve their phishing detection, we next evaluate which in-game interaction mechanics introduced in *Phooled!*, as detailed in Chapter 4, might have contributed to training our study participants.

5.3 Evaluation of *Phooled!* In-Game Interactions

In this section, we evaluate the in-game interaction mechanics of *Phooled!* using in-game telemetry data to understand how gameplay may have contributed to the IES improvement of non-Control participants who played *Phooled!* ($n = 52$). We first examine whether Collaborative training during the Group Pick round facilitated knowledge transfer of phishing cue detection (Section 5.3.1), then whether player roles affected classification performance (Section 5.3.2), whether feedback engagement contributed to improvement (Section 5.3.3), and finally whether in-game chat activity played a role (Section 5.3.4).

5.3.1 Effect on Cue Detection

As described in Chapter 4, the Group Pick round was designed as a collaborative round, where groups worked collectively to identify phishing cues in the email samples. We examine whether this collaboration helped knowledge transfer of phishing cue identification

ability. As noted at the beginning of this chapter, the 28 Collaborative participants were batched into 10 groups of 2–5 players.

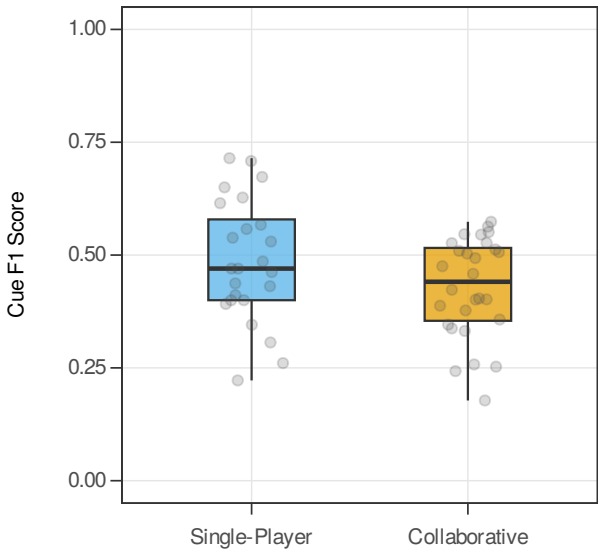


Figure 5.16: Boxplot of cue F1 scores by treatment condition

Collaborative groups ($n = 10$, $M = 4.03$, $SD = 0.74$) selected fewer cues per email than Single-Player participants ($n = 24$, $M = 5.13$, $SD = 1.67$), a significant difference ($t(32) = -3.00$, $p = .01$, $d = 0.75$). To measure group consensus on these selections, Jaccard similarity was calculated for email samples within each group. Groups agreed on 57% of their cue selections ($J = 0.57$), while the similarity for Single-Player participants was 38%. A one-tailed Wilcoxon rank-sum test confirmed that Collaborative groups had significantly higher agreement ($W = 372$, $p = .02$, $r = .25$).

To examine group influence on cue selection, the relationship between subjective norms and cue similarity was measured. SN1 showed no relationship with cue similarity ($\rho = 0.04$, $p = .80$). In contrast, SN2 showed a significant negative correlation ($\rho = -0.44$, $p = .030$), suggesting that participants who reported higher SN2 scale values showed less conformity to their groups. In other words, participants who had high SN2 values (i.e. influence of authority-based subjective norms) conformed less to their group's cue selections.

Figure 5.16 shows the distribution of cue detection F1 scores, with normality diagnostics in Appendix C.10. Notably, any collaborative benefit during cue detection is absent in the figure, which shows no visible difference in cue detection performance across treatment conditions. Despite higher consensus, Collaborative and Single-Player participants did not differ significantly in their F1 scores ($W = 414$, $p = .20$). To measure cue detection improvement over the course of training, a beta regression GLMM was fitted predicting F1 scores based on the number of cards seen, treatment condition, and sample type, while controlling for pre-training IES ($n = 350$ observations). A beta regression was used as F1 scores are bounded between 0 and 1. The model explained 81.2% of total variance (conditional $R^2 = .812$), of which 46.3% was due to fixed effects alone (marginal $R^2 = .463$).

The fixed effect of treatment on F1 scores was not statistically significant, though a marginal trend was observed ($\chi^2(1) = 3.65$, $p = .056$). However, F1 scores improved significantly as participants saw more cards during the course of the game ($\chi^2(1) = 12.82$, $p < .001$). The effect of treatment condition on F1 scores differed by sample type ($\chi^2(1) = 4.58$, $p = .032$). Participants improved their F1 scores faster for Safe samples ($b \approx 0.34$) than for Phishing samples ($b \approx 0.12$), suggesting that they learned to improve their cue detection in Safe emails more quickly during the game compared to Phishing emails. Although Collaborative groups showed higher consensus in their cue selections, both treatments improved their cue detection at similar rates, with no significant advantage from Collaborative play.

5.3.2 Effect of Game Roles on Classification Performance

As described in Chapter 4, the Group Pick round assigned participants one of three roles: Trickster, Defender, or Spectator. We examine whether these roles affected in-game classification performance. Figure 5.17 shows the distribution of IES values by game role, with normality diagnostics in Appendix C.9. Tricksters had the lowest IES values ($M = 9.69$, $SD = 7.44$), followed by Spectators ($M = 19.94$, $SD = 14.36$), and finally Defenders ($M = 22.43$, $SD = 13.14$).

A Gamma GLMM was fitted predicting IES from game role and sample type, while

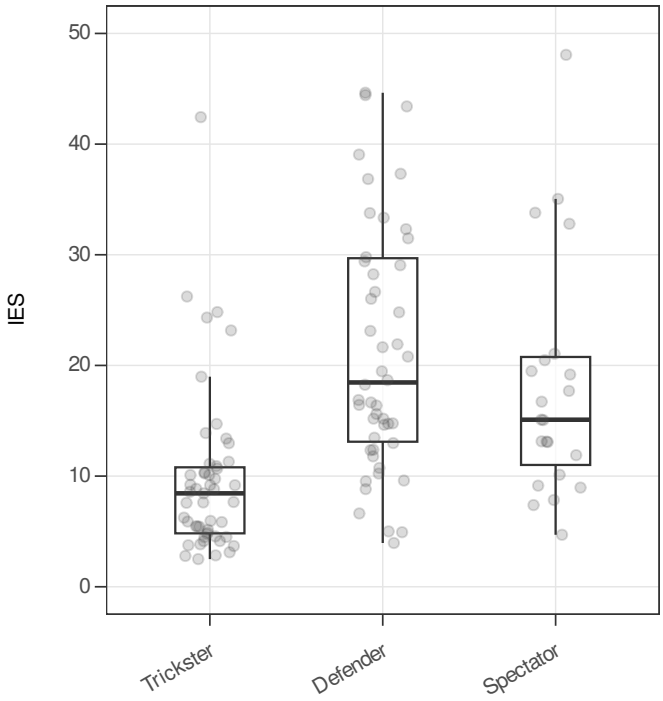


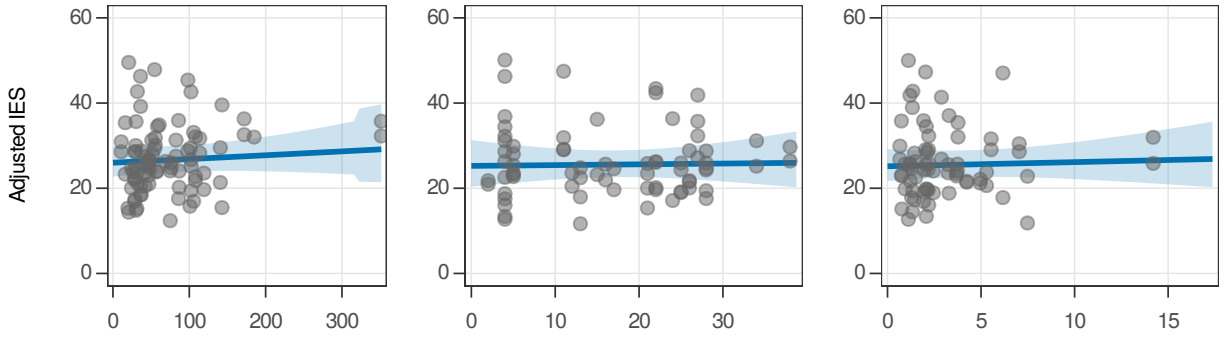
Figure 5.17: Boxplot of IES values by game role

controlling for cohort ($n = 291$ observations). The model explained 36.2% of total variance (conditional $R^2 = .362$), of which 26.2% was due to fixed effects alone (marginal $R^2 = .262$). The fixed effect of game role on IES values was statistically significant ($\chi^2(3) = 58.83$, $p < .001$).

Since Tricksters classified the same card a second time after being exposed to it already in the Quick Pick round, their lower IES values reflect repeated exposure. Critically, the IES values of the Defender and Spectator roles did not differ from each other ($z = 0.70$, $p = .896$). This indicates that the passive Spectator role had similar IES values to the active Defender role. Therefore, players in passive roles showed similar game engagement as players in active roles.

5.3.3 Effect of Feedback Engagement

Phooled! provides feedback on classification and cue detection correctness through two mechanisms: in-game feedback after the Classification and Find the Tricks stages, and a dedicated Feedback and Review round at the end. As described in Section 4.1.2, providing feedback immediately after each game stage was designed to ensure it is close to the learning activity, in line with the learning theory principles discussed in Chapter 2. The dedicated Feedback and Review round at the end provides additional time for players to review and reflect on their learning without time pressure. We examine whether greater engagement with this feedback was associated with better phishing detection after gameplay.



(a) Feedback time (s) (b) Cards reviewed (c) Time per card (s)

Figure 5.18: Feedback engagement and adjusted post-training IES

We first calculated the total time spent on feedback review within the game. This metric consisted of time spent in the feedback round and the review stages after the Classification stage and Find the Tricks stage. Participants spent a median of 55.41 seconds (IQR = 70.20) on total feedback review, indicating large variation in time spent learning from feedback.

Figure 5.18(a) shows the relationship between adjusted IES and total feedback time. A Gamma GLMM was fitted predicting IES from total feedback time, while controlling for pre-training IES, cohort, and treatment condition ($n = 92$ observations). The model explained 66.0% of total variance (conditional $R^2 = .660$), of which 42.5% was due to fixed

effects alone (marginal $R^2 = .425$). The fixed effect of total feedback time on IES was not statistically significant ($\chi^2(1) = 0.64$, $p = .424$).

Feedback metrics were computed based on the number of samples reviewed (Figure 5.18(b)) and the median duration per sample (Figure 5.18(c)). Participants reviewed a median of 19 ($IQR = 21$) samples, with a median review time of 2.18 seconds ($IQR = 2.74$) per sample, suggesting that most participants did not engage deeply with the feedback. A Gamma GLMM was fitted predicting IES from both the number of samples reviewed and the median time per sample, while controlling for pre-training IES, cohort, and treatment condition ($n = 76$ observations).

Neither fixed effect, the number of samples reviewed ($\chi^2(1) = 0.02$, $p = .890$) nor the median time per sample ($\chi^2(1) = 0.14$, $p = .705$), was statistically significant. The model explained 65.3% of total variance (conditional $R^2 = .653$), of which 42.2% was due to fixed effects alone (marginal $R^2 = .422$). In conclusion, feedback engagement did not predict phishing detection improvement, perhaps due to a general lack of engagement of participants with these two feedback mechanisms.

5.3.4 Effect of Game Chat Activity

The in-game chat allowed Collaborative participants to communicate with each other during gameplay. We examined whether phishing task discussion during Collaborative gameplay predicted phishing detection improvement. Across all cohorts, participants sent a total of 277 messages over 20 game sessions. The chat messages from these sessions were mined and manually classified using an iterative manual data coding process with repeated bucketing, forming two main categories of messages. Figure 5.19 shows these categorised messages per player over the course of training, as detailed below:

- **On-topic messages**

    1. *Phishing Task Discussion*: messages that discussed different aspects of the phishing detection and learning task, e.g., “but that wasnt the end goal of the email”, “eh, not sure about the link mismatch”.

2. *Game Discussion*: discussions related to in-game interaction mechanics and objectives within the game that showed engagement with the learning activity, e.g., "i have terrible cards", "why am i the only one with 2 hearts lol what did i do wrong".

- **Off-topic messages**

1. *Game Coordination*: messages related to game setup and management of the learning activity itself, e.g., "waiting on you to start", "maybe his/her browser crashed..".
2. *Emotional Expressions*: e.g., "let's go!", "weee".
3. *Greetings*: e.g., "Hello :-)", "hi guys".
4. *Well-wishes*: e.g., "good luck", "we got this".
5. *Humour*: e.g., "womp womp my friend", "who volunteers as a tribute?".

Messages that contained dual intentions, such as "Its a fish", were classified giving higher priority to on-topic sub-categories. It was observed that chat message activity varied across cohorts. IT Professionals hardly sent any messages ($M = 0.78$, $SD = 0.97$), none of which were classified as on-topic.

Remote participants sent an average of 9.83 messages ($SD = 10.86$), of which 4.17 were on-topic ($SD = 6.98$). HS Students were the most active in the chat, sending an average of 18.63 messages ($SD = 25.95$), of which 9.25 were on-topic ($SD = 14.52$). As shown in Figure 5.19, HS Students chatted continuously throughout the game with messages spread across the full game duration. In contrast, messages from Remote participants were more sparse and appeared to be used more for game coordination. The IT Professionals cohort was dropped from further analysis due to the absence of on-topic messages and a small sample size.

Figure 5.20 shows that Phishing Task Discussion messages correlated significantly with both in-game classification score ($\rho = 0.59$, $p = .006$) and cue identification accuracy ($\rho = 0.58$, $p = .009$). To test whether these messages also predicted phishing detection

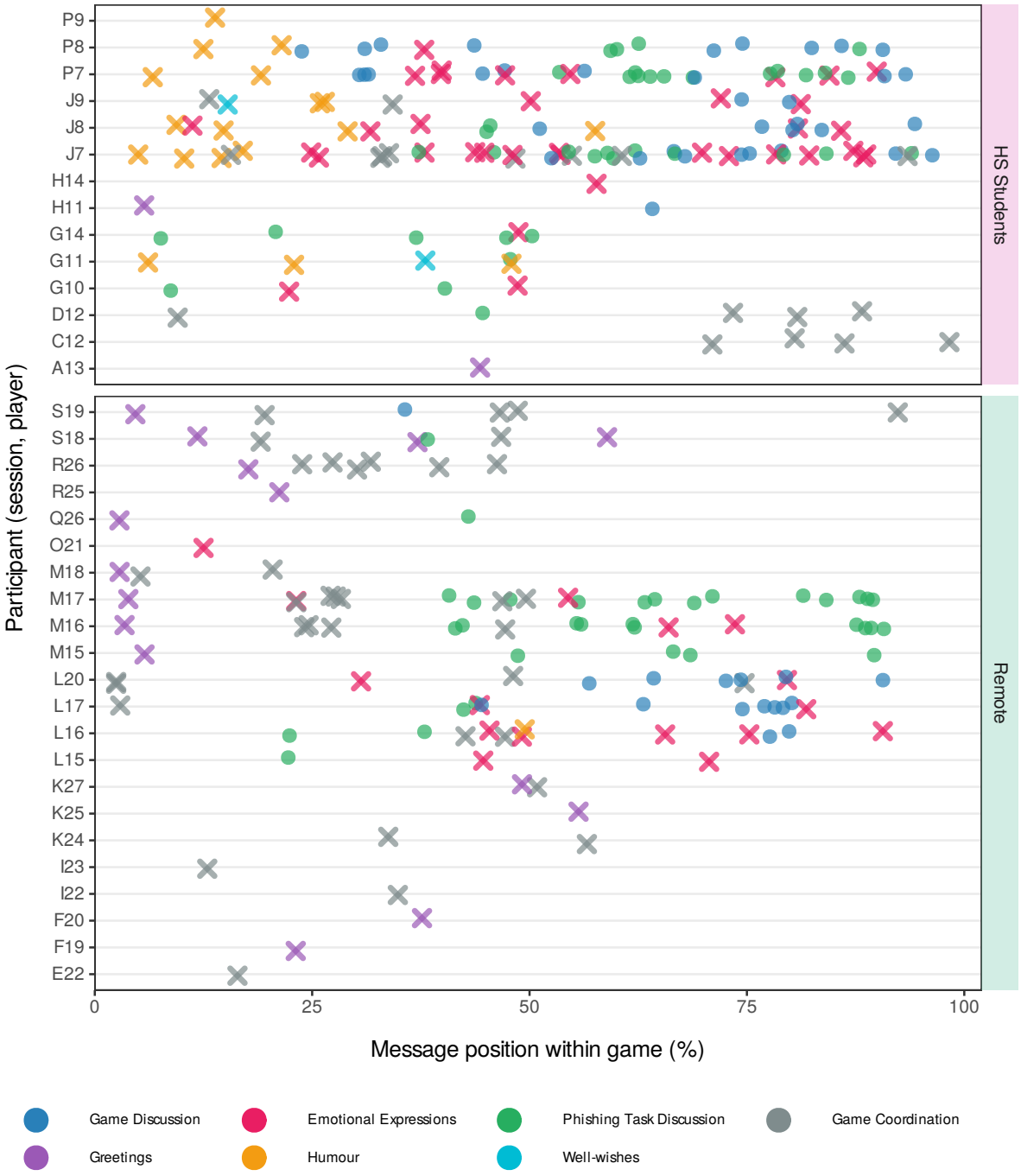
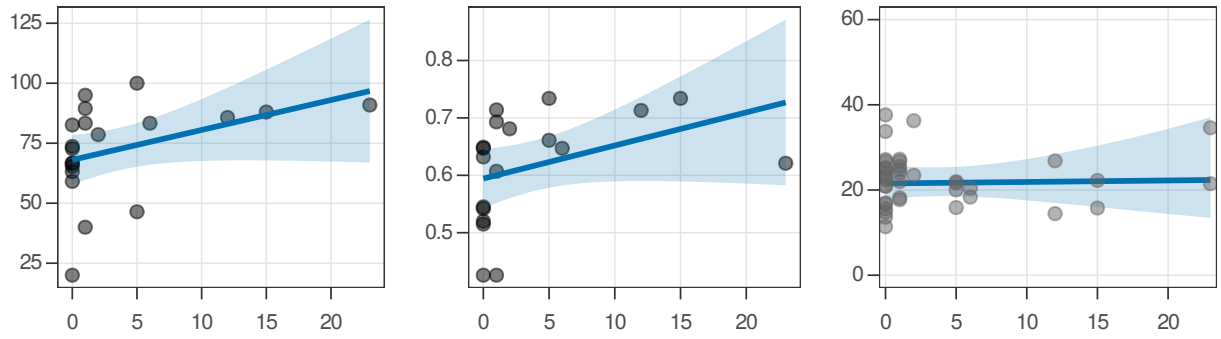


Figure 5.19: Timeline of chat messages for HS Students and Remote participants



(a) Score (b) Accuracy (c) Adjusted IES

Figure 5.20: Phishing Task Discussion messages (on the x axis) vs. in-game score, cue accuracy, and post-training IES

improvement after training, a Gamma GLMM was fitted predicting IES from message count, while controlling for pre-training IES and cohort ($n = 20$ participants). The model explained 68.3% of total variance (conditional $R^2 = .683$), of which 44.4% was due to fixed effects alone (marginal $R^2 = .444$). The fixed effect of Phishing Task Discussion messages on IES was not statistically significant ($\chi^2(1) = 0.02$, $p = .901$), as shown in Figure 5.20c.

The fixed effect of Cohort, on the other hand, was significant ($\chi^2(1) = 9.82$, $p = .002$), with HS Students having lower IES values than Remote participants ($b = 0.51$, $p = .002$). Therefore, while Phishing Task Discussion correlated with better in-game performance, this did not translate to post-training improvement. HS Students both chatted and improved more than the Remote cohort, but the apparent relationship was driven by cohort-specific factors rather than chat activity itself.

5.3.5 Summary of In-Game Interaction Mechanics Evaluation

In this section, we evaluated in-game interaction mechanics as determinants of improving phishing detection efficiency after treatment. Cue detection F1 scores improved significantly as participants encountered more email samples, regardless of treatment condition. The repeated exposure to email samples through the structure of the game rounds and roles benefited learning inside the game, as demonstrated by the increased IES scores for

the Trickster role. The passive Spectator role performed just as well as the active Defender role during the game, suggesting that passive roles can keep users engaged in the learning even when they are not the focus of the round. When participants worked together in groups, their consensus on cue selections was higher, indicating that the Find the Tricks stage facilitated group agreement on phishing detection.

However, these individual in-game interaction mechanics could not be linked to improved post-training IES scores. Though *Phooled!* was designed to provide feedback through both immediate in-game feedback and a dedicated review round, in line with the learning theory principles discussed in Chapter 2, low engagement with these mechanisms led to no measurable benefit. This suggests that feedback mechanisms alone are insufficient without active learner engagement. Features like in-game chat correlated with better in-game performance but did not translate to improvement in detection after training. Having evaluated the in-game interaction mechanics of *Phooled!*, we then investigate the determinants of phishing email detection difficulty within the game itself.

5.4 Determinants of Phishing Email Detection Difficulty

In this section, we examine factors that influenced the difficulty of phishing email detection within *Phooled!*. Using in-game telemetry data, we analyse how email sample type and phishing cues affected detection performance in Section 5.4.2. We also assess the effect of time pressure on classification correctness (Section 5.4.1), whether participants' information processing style predicted in-game performance (Section 5.4.3), and whether participants were good judges of their own classification correctness (Section 5.4.4).

5.4.1 Effect of Time Pressure on Classification Correctness

Figure 5.21 shows the distribution of classification duration across the Quick Pick round and the Classification stage. Classification duration was shorter in the Quick Pick round ($M =$

12.20 s, $SD = 8.95$) than in the Classification stage ($M = 15.93$ s, $SD = 7.58$), consistent with the game design intention to push more heuristic behaviour in the Quick Pick round by introducing more time pressure. Q–Q plots and Shapiro-Wilk tests indicated significant deviations from normality for both the Quick Pick round ($p < .001$) and the Classification stage ($p < .001$), as shown in Appendix C.8. Consequently, a paired Wilcoxon signed-rank test was used to compare duration across these two game states on a per-participant basis ($n = 53$). Classification duration was significantly longer in the Classification stage than in the Quick Pick round ($V = 255$, $p < .001$, $r = 0.56$), with a median increase of 3.04 s ($IQR = 7.65$).

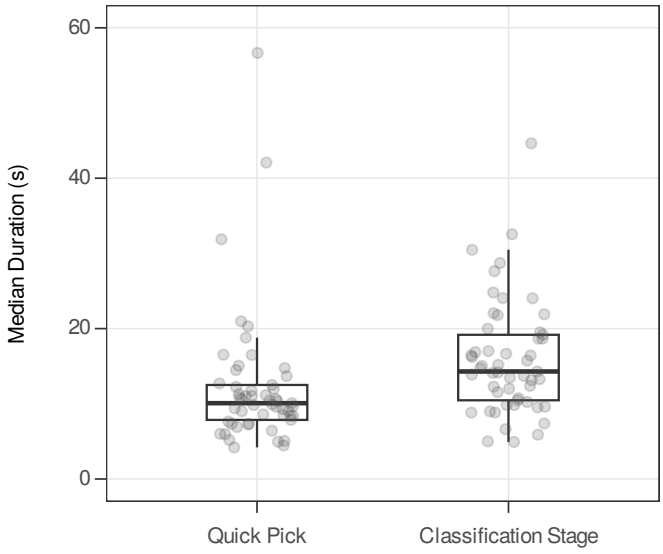


Figure 5.21: Distribution of median classification duration across Quick Pick round and Classification stage

Despite this increased classification duration in the Classification stage, the increased time spent for classification did not lead to better outcomes. A binomial GLMM was fitted predicting classification accuracy from classification duration, while controlling for sample type, cohort, and treatment ($n = 1{,}112$ observations), with DHARMa residual diagnostics confirming adequate model fit. The model explained 30.5% of total variance (conditional $R^2 = .305$), of which 13.6% was due to fixed effects alone (marginal $R^2 = .136$).

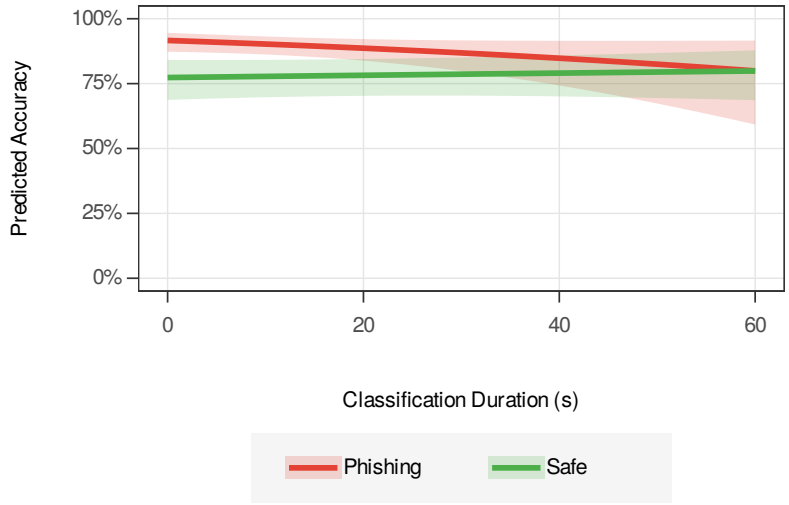


Figure 5.22: Predicted classification accuracy by duration and sample type (95% CI bands)

The fixed effect of classification duration on accuracy was not statistically significant ($\chi^2(1) = 0.07$, $p = .791$). As shown in Figure 5.22, the accuracy of Safe samples appears to increase with duration while the accuracy of Phishing samples slightly declines, but this interaction was also not significant ($\chi^2(1) = 2.98$, $p = .084$). This suggests a trend where additional processing time benefits the classification of Safe emails, while the same may be harmful for Phishing emails. However, neither trend reached statistical significance. Taken together, while participants took longer to classify in the Classification stage, this additional time did not translate to improved classification accuracy.

5.4.2 Effect of Email Sample Type and Cues

We next examine which properties of an email influenced classification difficulty, focusing on email sample type (either Phishing or Safe) and the phishing cues listed in Section 4.4.2.

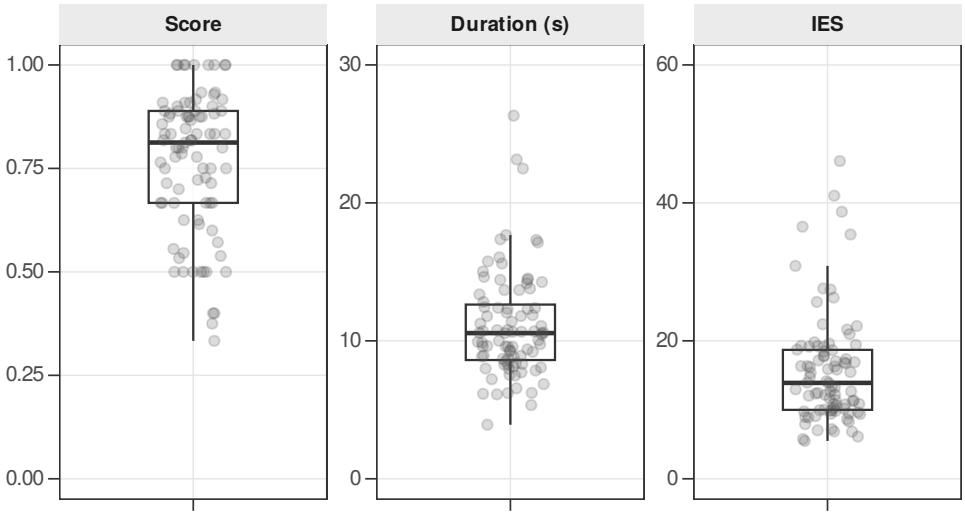


Figure 5.23: Boxplots of email sample classification score, duration, and IES

Sample Type

To examine phishing detection performance per email sample, three aggregated metrics were computed: classification score, classification duration, and IES. The boxplot of these distributions is shown in Figure 5.23. This figure shows the distribution of user classification performance across the email samples for these three measures. Across 87 samples, the averages of these metrics were: classification score ($M = 0.77$, $SD = 0.17$), duration ($M = 11.09$ s, $SD = 3.85$), and IES ($M = 15.70$, $SD = 8.01$). A classification score of 0.77 indicates that participants performed well above the chance level of 0.50.

The Q–Q normality plots for these metrics are shown in Appendix C.5. Visual inspection and Shapiro-Wilk tests indicated significant deviations from normality for all three distributions ($p < .001$). These deviations arose from a subset of samples having low classification accuracy and long classification duration, which resulted in higher IES values. The relationship between classification duration and accuracy was examined in the previous subsection.

To further examine the effect of sample type on sample difficulty, IES was compared

between Safe and Phishing sample types. The distribution of IES by sample type is shown in Figure 5.24. Safe samples had a higher mean IES ($M = 21.30$, $SD = 7.68$) compared to Phishing samples ($M = 13.00$, $SD = 6.73$), indicating that Safe samples were more difficult to classify. As shown in Appendix C.6, Q–Q plots indicated deviations from normality for both distributions. These deviations were confirmed by Shapiro-Wilk tests for Phishing ($W = 0.71$, $p < .001$) and Safe ($W = 0.92$, $p = .038$) samples. A Gamma GLM was fitted predicting IES from sample type ($n = 87$ samples). The model accounted for 28% of the variance in the data ($R^2 = .280$). The fixed effect of sample type on IES was statistically significant ($z = 5.50$, $p < .001$), with Safe samples having 64% higher IES values than Phishing samples.

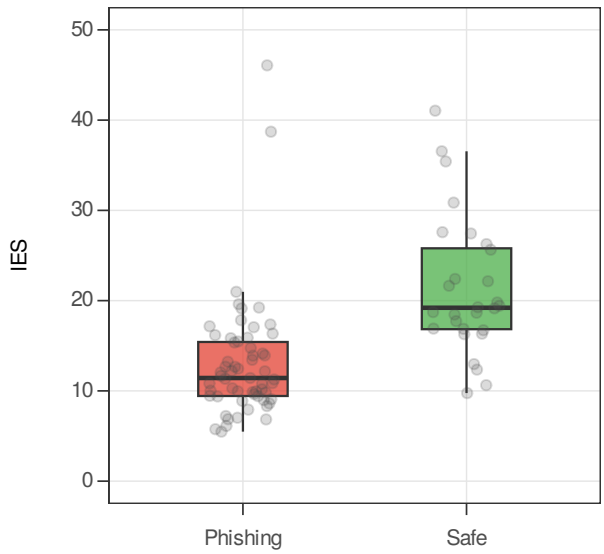


Figure 5.24: Boxplot of IES by email sample type

Cue Identification

To evaluate participant performance on phishing cues, in-game cue selections were compared against the ground truth in the sample bank to compute precision, recall, and F1 score for each cue ($n = 350$ observations). A low F1 score for a cue indicates higher false

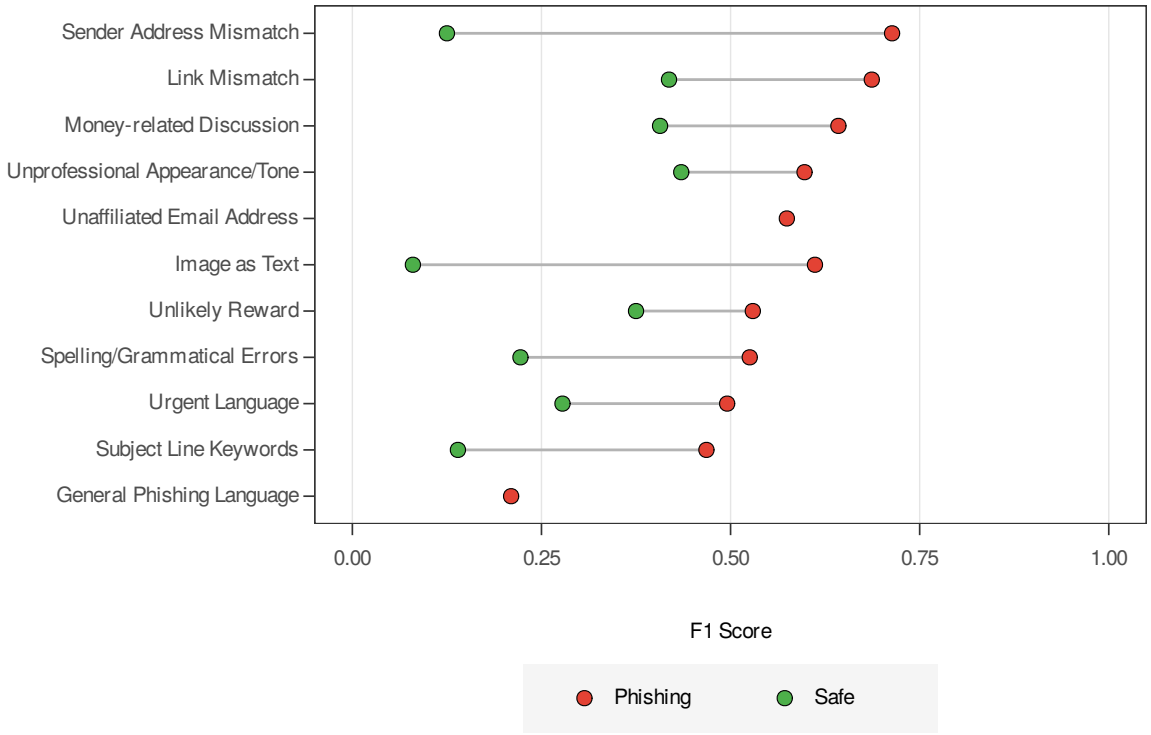


Figure 5.25: F1 scores for phishing cue identification by sample type

positive or false negative rates. In the case of perfect cue identification, the F1 score would be 1.0.

Across all cues, F1 scores ranged from 0.69 (Sender Address Mismatch) to 0.18 (General Phishing Language), with lower F1 score values indicating worse performance. The F1 scores were affected by decreases in either precision or recall, i.e., increases in FPR or FNR respectively. The General Phishing Language cue had a precision of 0.11, which was the lowest among all cues. In contrast, the Spelling/Grammatical Errors cue had the lowest recall at 0.45, followed by Link Mismatch (recall = 0.52) and Sender Address Mismatch (recall = 0.59).

When analysed by sample type (Figure 5.25), F1 scores were higher for Phishing samples than for Safe samples across all measured cues. F1 scores for Phishing samples were consistent with the overall F1 scores, while Safe samples had lower F1 scores across all

cues. The cue with the highest F1 score among Phishing samples was Sender Address Mismatch ($F1 = 0.71$), while for Safe samples, it was Unprofessional Appearance/Tone ($F1 = 0.44$). The lower F1 scores for Safe samples suggest that participants found it more "tricky" to identify cues in Safe emails, or falsely identified cues that were not present in them. The distribution of IES for select cues is shown in Figure 5.26. To examine whether the presence of an individual cue predicted classification difficulty, the model from Section 5.4.2 was extended to include sample cues as shown below:.

$$\text{IES}_i \sim \text{SampleType}_i + \sum_{k=1}^{K} \text{SampleCue}_{ik} \tag{5.3}$$

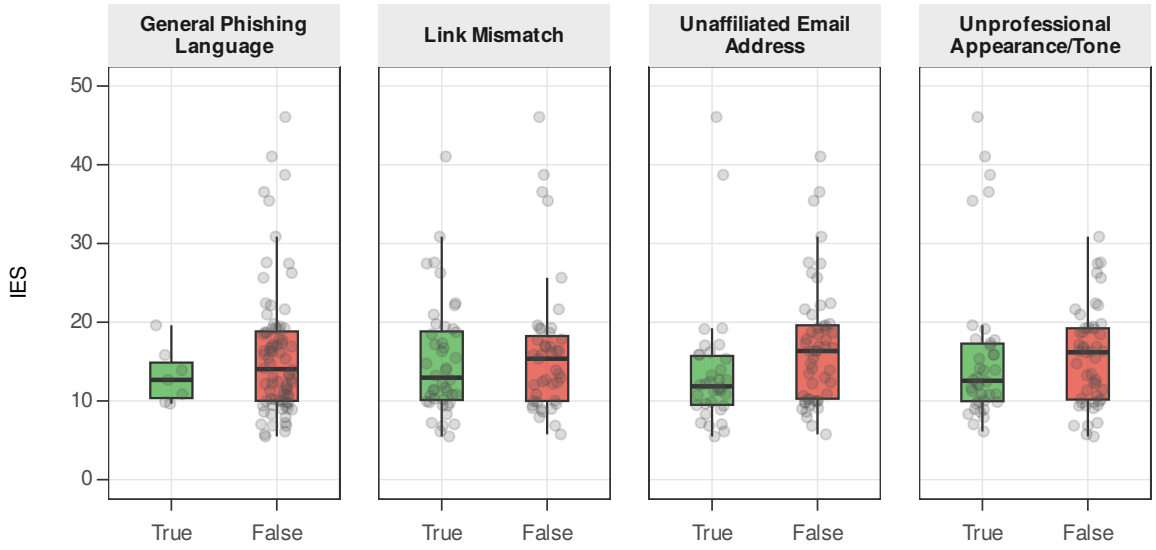


Figure 5.26: Boxplots of IES for select email sample cues

Sample type remained a significant predictor of IES in the extended model ($z = 2.91$, $p = .004$), but none of the individual sample cues were significantly associated with IES after controlling for sample type ($n = 87$ samples). It is likely that participants processed emails holistically in a heuristic manner, rather than systematically checking individual indicators. They operated on a perceived notion of the overall "trickiness" of an email,

rather than focusing on specific cues.

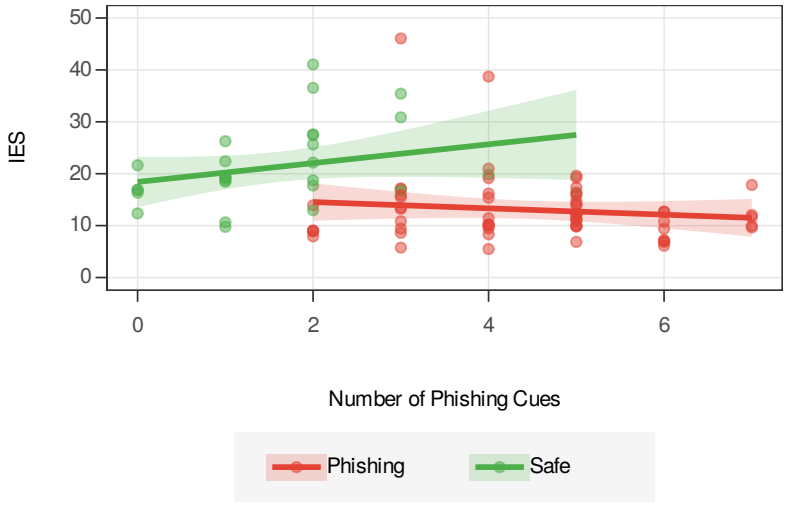


Figure 5.27: Sample IES by cue count and sample type

To examine whether cues had an overall effect, a cue count metric was calculated as the total number of cues present in each sample. Phishing samples contained significantly more cues ($n = 59$, $M = 4.47$, $SD = 1.39$) than Safe samples ($n = 28$, $M = 1.61$, $SD = 1.23$). The relationship between cue count and IES is shown in Figure 5.27, displaying different trends for Phishing and Safe emails.

A Gamma GLM was fitted to quantify this relationship, while controlling for sample type ($n = 87$ samples). The interaction between sample type and cue count was significant ($b = 0.15$, $z = 1.97$, $p = .048$). In Phishing emails, more cues were associated with lower IES values ($b = -0.05$), while in Safe emails, more cues were associated with higher IES values ($b = 0.10$). This effect corresponds to approximately 5% lower IES per additional email cue in Phishing emails, and approximately 10% higher IES per additional email cue in Safe emails.

Mining Effective Phishing Cues

To further investigate which email features influenced classification difficulty, structural and semantic features were derived from the email sample bank as follows:

- **Subject features:**
  - *Subject line length*: Number of characters in the subject line after removing placeholders and trimming whitespace.
  - *Reply indicator*: Emails whose subject lines begin with reply prefixes (e.g., "Re:"). For example, "Re: BTC wallet details".
  - *Urgent subject*: Subject lines containing urgency keywords (e.g., *urgent*, *alert*), exclamation marks, or all-caps text. For example, "Act fast! Your latest surveys are in".
- **Sender features:**
  - *Institutional domain*: Indicates whether the email originated from a domain that is not open to public sign-ups (i.e., not from free email providers such as Gmail, Outlook, Yahoo, or iCloud).
  - *From Subdomain*: Indicates whether the sender email address contains a subdomain (e.g., "FirstnameLastname@cmail.carleton.ca").

A boxplot of the distribution of IES for select sample features is shown in Figure 5.28. A Gamma GLM was fitted to examine the relationship between these sample features and IES, while controlling for sample type ($n = 87$ samples). Emails originating from an institutional domain had significantly lower IES values than those from public email providers like Gmail ($b = -0.48$, $z = -3.52$, $p < .001$), with non-institutional samples having 62% higher IES values.

Email samples with senders using subdomains in their From Address showed a marginal but non-significant positive association with IES ($b = 0.15$, $z = 1.72$, $p = .086$). None of the subject features had a significant association with IES. From the features examined, only the sender's domain significantly influenced classification difficulty.

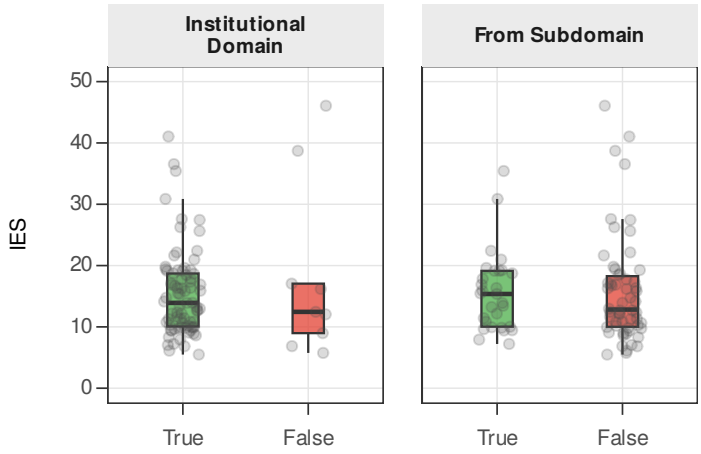


Figure 5.28: Boxplots of IES for select email sample features

5.4.3 Effect of Information Processing Style

In Section 5.2, we briefly noted that the Heuristic and Systematic processing scales did not individually predict phishing detection efficiency after training. In Section 5.4.1, we found that classification duration did not predict accuracy. We explore this finding here by examining whether participants' Heuristic and Systematic processing scores predicted their classification efficiency differently across the Quick Pick round and the Classification stage.

Figure 5.29 shows the distribution of Heuristic and Systematic processing scores by cohort, with Q–Q normality plots in Appendix C.11. The Heuristic processing scores were similar across cohorts: HS Students ($M = 3.00$, $SD = 0.85$), IT Professionals ($M = 3.05$, $SD = 1.06$), and Remote ($M = 2.84$, $SD = 0.86$). However, IT Professionals were the most Systematic ($M = 4.33$, $SD = 0.64$), followed by Remote ($M = 4.01$, $SD = 1.08$) and HS Students ($M = 3.76$, $SD = 0.90$). A Spearman correlation test found that Heuristic processing was not correlated with either Quick Pick IES ($\rho = 0.04$, $p = .80$) or the number of cards classified during Quick Pick ($\rho = -0.08$, $p = .60$), indicating that Heuristic processing did not lead to better efficiency in Quick Pick, either in terms of duration of classification or number of emails classified.

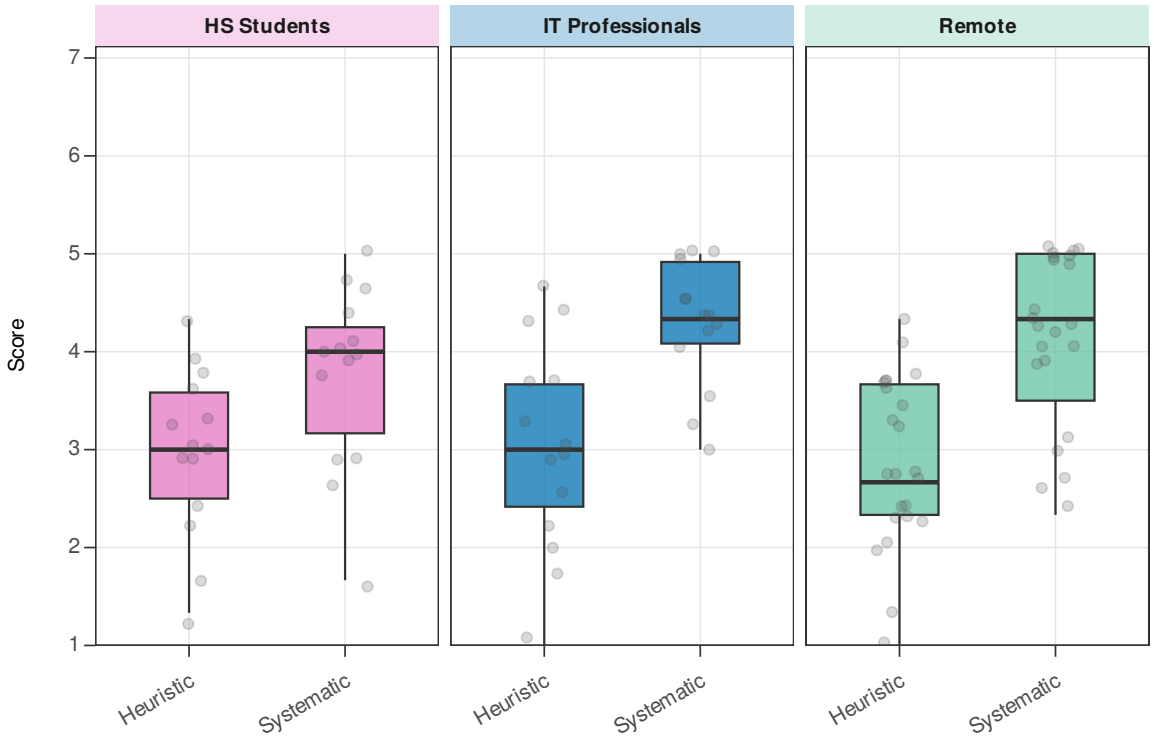


Figure 5.29: Boxplot of Heuristic and Systematic processing scores by cohort

In contrast, the Systematic processing scale showed a moderate negative correlation with the Classification stage IES ($\rho = -0.30$, $p = .030$). A Gamma GLM supported this relationship, showing a marginal fixed effect of Systematic processing ($\chi^2(1) = 3.69$, $p = .055$), with a significant cohort interaction ($\chi^2(2) = 6.41$, $p = .040$). This relationship is shown in Figure 5.30.

We can observe that the fixed effect was largely driven by IT Professionals ($b = -0.58$, CI $[-1.01, -0.15]$), for whom higher Systematic scale values predicted better efficiency during the Classification stage. This was not true for HS Students ($b = 0.11$, CI $[-0.22, 0.43]$) nor Remote participants ($b = -0.14$, CI $[-0.34, 0.05]$). The fixed effect among IT Professionals differed significantly from HS Students ($z = 2.54$, $p = .039$).

Thus, we note that neither self-reported Heuristic nor Systematic processing translated

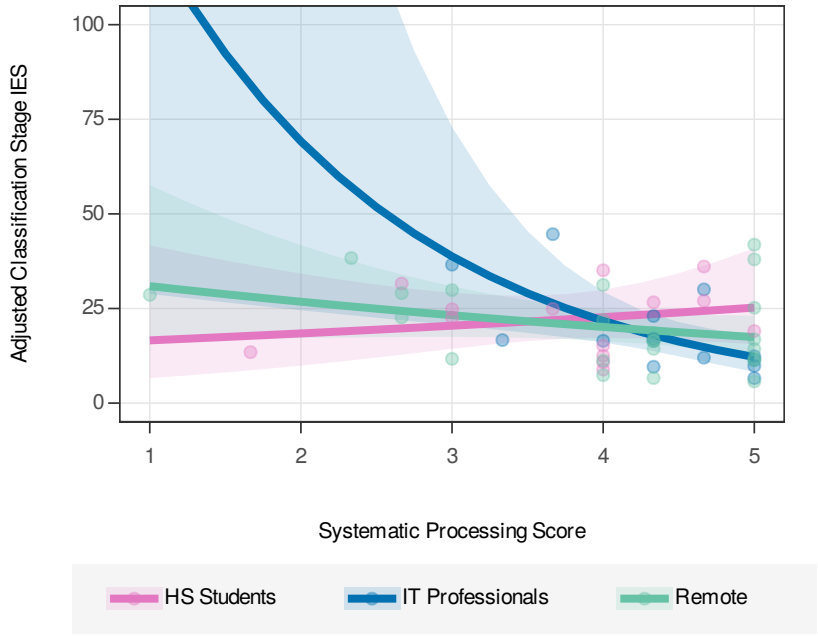


Figure 5.30: Adjusted Classification Stage IES by Systematic processing score and cohort (95% CI bands)

to better in-game performance. In fact, most participants defaulted to Heuristic processing, with Systematic processing only benefiting cohorts like IT Professionals who are more likely to have domain knowledge on the task at hand. This further reinforces our finding from Section 5.4.2, that most participants processed emails heuristically and holistically, with IT Professionals as the exception.

5.4.4 Perceived Correctness of Classification

To investigate participants' card preference during the Trick Pick and Group Pick rounds, two aggregated metrics were calculated for each card: the selected ratio and the played ratio. A card is considered selected if it was kept by the user during the Trick Pick round. It is considered played if it was chosen by the Trickster to use against the Defender during the Group Pick round.

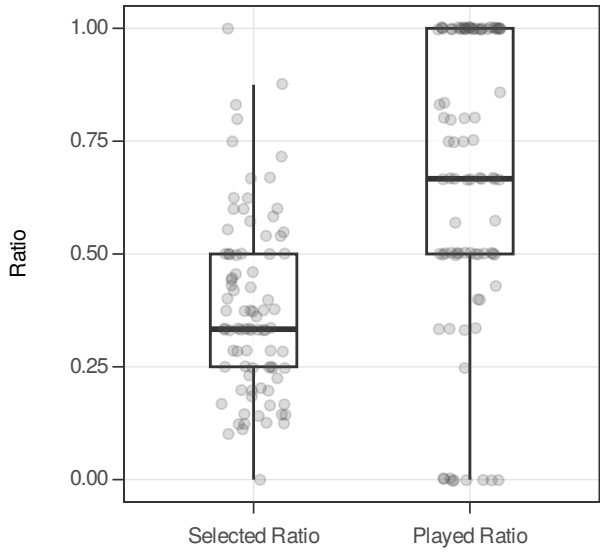


Figure 5.31: Boxplots of card selection and play ratios per email sample

The distributions of these ratios are shown in Figure 5.31. The mean selected ratio ($M = 0.38$, $SD = 0.20$) was lower than the mean played ratio ($M = 0.67$, $SD = 0.32$), suggesting that participants tended to play most of their cards once selected. Shapiro-Wilk tests for both selected ratio ($p = .010$) and played ratio ($p < .001$) confirmed deviation from normality, consistent with the Q–Q plot shown in Appendix C.7.

To assess the influence of sample difficulty on card preference, classification correctness and classification duration during the Quick Pick round are shown in Figures 5.32 and 5.33, respectively. Binomial GLMMs were fitted predicting card selection from classification correctness and classification duration, while controlling for sample type, cohort, and treatment ($n = 697$ observations). The fixed effect of classification correctness on card selection was not statistically significant ($z = 0.51$, $p = .612$), whereas classification duration was significant ($z = 4.66$, $p < .001$). Each additional second spent classifying an email during the Quick Pick round increased the odds of selection during the Trick Pick round by 4%. Safe samples had higher odds of being selected than Phishing samples ($z = 2.26$, $p = .024$), with 52% higher selection odds, despite accounting for less than one third of all cards in the participants' collective hand ($n = 197$ vs. $n = 500$).

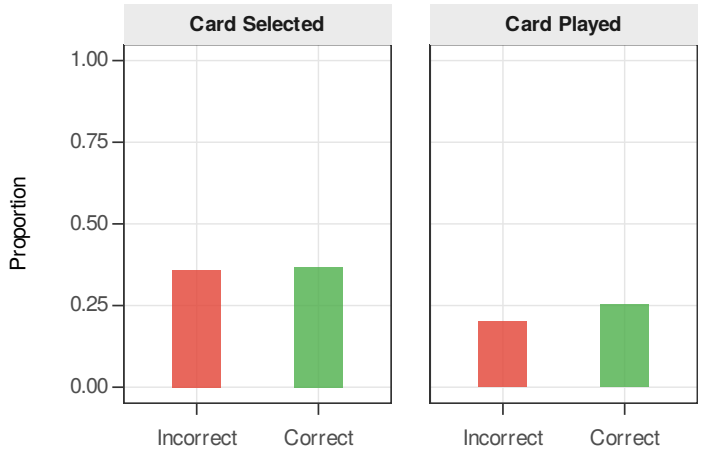


Figure 5.32: Card selection and play proportions by classification correctness (95% CIs)

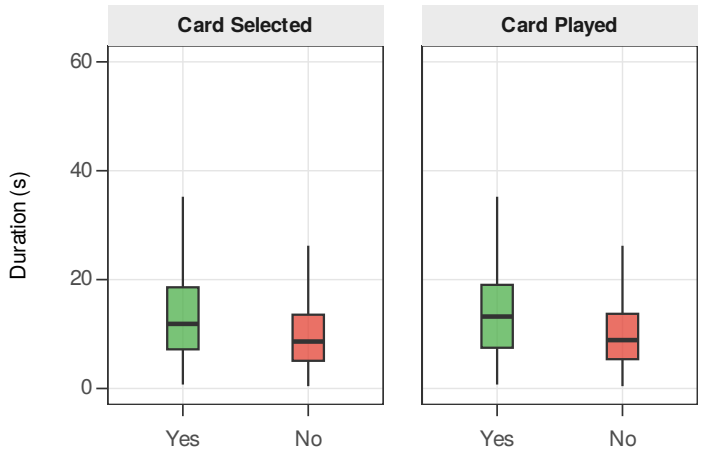


Figure 5.33: Boxplot of classification duration by card selection

In the Group Pick round, classification correctness showed a marginal association with card play ($z = 1.85$, $p = .065$), while duration remained a significant predictor ($z = 4.87$, $p < .001$), with each additional second increasing the odds of play by 5%. Similarly, Safe samples had 81% higher odds of being played than Phishing samples ($z = 3.01$, $p = .003$). Participants preferentially selected and played cards they spent longer classifying, with a consistent preference for Safe samples across both rounds. However, these choices were not

based on actual classification correctness.

5.4.5 Summary of Phishing Email Detection Difficulty

Overall, we found that Safe emails were consistently harder to detect than Phishing emails in our study. Therefore, the primary determinant of detection difficulty was sample type itself. Cues in email samples did not independently predict difficulty, but their overall presence or count had opposite effects, with Safe samples growing in difficulty as more cues were present. Among email features, the sender's domain was the main factor in determining detection difficulty, with public domain names "tricking" more users. We also observed that participants' detection accuracy did not improve as they took more time on a sample.

Participants preferentially engaged with cards they had spent more time classifying, not necessarily ones they were confident in. This suggests that participants processed emails in a heuristic manner, relying on a general notion of "trickiness" that triggered detection or avoidance behaviour rather than focusing on specific cues. We found that self-reported processing scales did not predict in-game performance, with the exception of IT Professionals who may have benefited from Systematic processing due to their domain knowledge or other inherent characteristics. Finally, participants have an inflated sense of their own phishing detection ability. This mismatch between actual correctness and perceived correctness may hinder their phishing detection performance during routine email processing.

5.5 Participant Feedback

Finally, we draw on our post-game feedback survey ($n = 50$ non-Control participants) and our field notes from the in-person focus groups to better understand study participants' outlook on *Phooled!*. We first present quantitative Likert scale responses from the survey in Section 5.5.1, followed by a thematic analysis of the remaining open-ended responses and field notes from focus group sessions in Section 5.5.2.

5.5.1 Quantitative Survey Responses

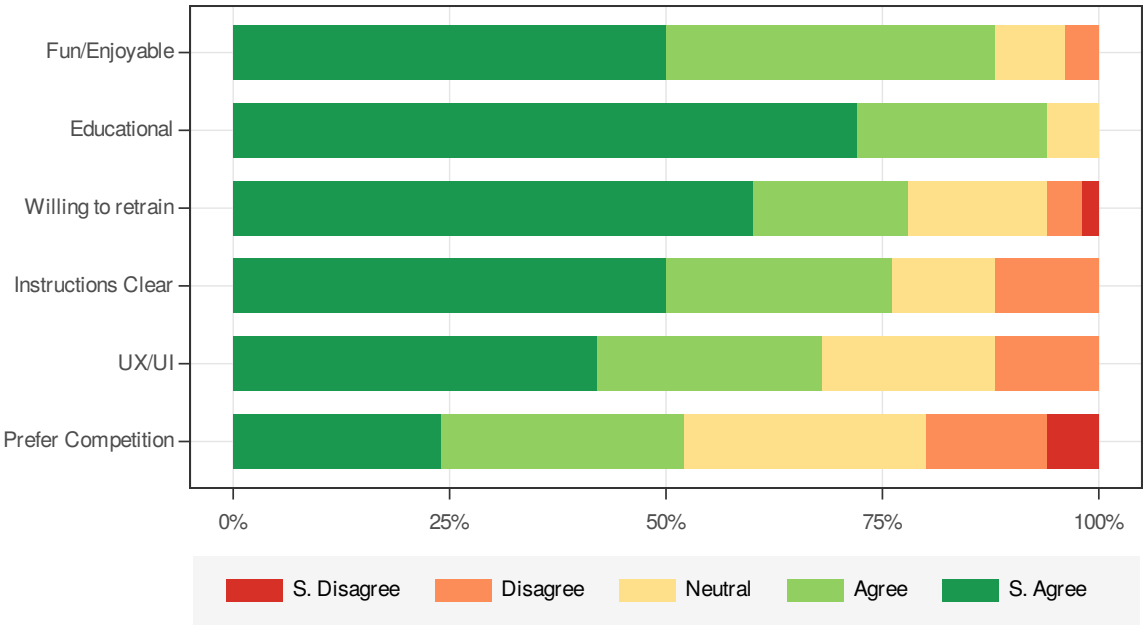


Figure 5.34: Stacked bar chart of post-game survey responses ($n = 50$)

Figure 5.34 shows the Likert scale responses from our six quantitative survey items. Likert scale scores showed that participants found the game to be fun and educational, with 88% rating it at least 4 out of 5 for "I found the Phooled game to be fun/enjoyable" and 94% for "I think the Phooled game is a good educational tool to help me learn how to avoid phishing emails". There were mixed responses to the game instructions and the user experience/interface (UX/UI). Around a third of participants did not find the interface intuitive and a quarter found that the game instructions were not clear.

Cohorts differed in their ratings, with Remote participants rating the highest values for 4 of the 5 items. HS Students were the least willing to retrain, with only 50% willing to do so, compared to approximately 90% of IT Professionals and Remote participants. Participants were evenly split on whether they preferred competitive or collaborative gameplay, suggesting that both competitive and collaborative elements are important in games.

5.5.2 Thematic Analysis of Open-Ended Survey Data and Focus Group Sessions

We examined the open-ended survey items and focus group notes to derive themes reflecting participant feedback. These themes were formed using thematic analysis [13]. A total of 30 codes were created from the participant feedback data and organised into five themes as detailed below.

User Feedback on Collaborative Learning

Participants in the Collaborative treatment valued the social aspect of *Phooled!*, one HS Student stating that the game "gave a cooperative way to interact, as you can work together and debate different ideas," while a Remote participant appreciated "the multiplayer aspect of it [...] as you can chat with other players." That said, not all participants advocated for collaborative elements. One IT Professional remarked, "How is that helping me? If I had more time... I won't see what other people did, [because] I only looked at [what] other people did," suggesting that collaboration distracted them from the learning task. This also provided an insight into why group consensus increased in the Find the Tricks stage without increasing accuracy as reported in Section 5.3.1. One individual suggested combining competition with collaboration, proposing that a "team vs teams [game] [...] [where] teams collab and try to beat the other team."

Perception of Phishing Detection Learning

Most participants felt that they learned more about phishing detection after playing *Phooled!*. During the focus group, most HS Students reported that the game helped them learn more about phishing emails. One HS Student stated, "I liked how while being a game, it is also an educational tool," while another Remote participant said that "the game was fun and taught me more about scam emails." In contrast, one IT Professional demanded more metadata for emails, stating that "[I] want [the] time and date that the email came [at] to compare [with] the time in the email," while another dismissed the game altogether,

stating "I [have taken] a lot of training, [and] I don't think I learnt more than [what I already knew]." This is consistent with the in-game performance analysis in Section 5.4.3, which suggested IT Professionals' deeper domain knowledge.

Feedback on Phooled! Game Design

Participants engaged positively with *Phooled!* in-game interaction mechanics. One Remote participant enjoyed "playing against the phishbot." The time pressure mechanism was evident as many players were concerned about "being rushed because of the ready [mechanism]."

Participants endorsed the gamification approach, with one Remote participant opining: "I think gamifying this subject is an awesome idea." The reward mechanism was particularly motivating to users, one engaged HS Student participant expressing dissatisfaction about the fact that they did "NOT [GAIN] MONEY WHEN I WAS CORRECT." Repetition was an issue, with one HS Student complaining that "I disliked how repetitive the middle of the game was," while another suggested that they would "love a section where you try and trick other players by designing your own scam emails."

Feedback on User Interface

Game instructions drew polarising opinions among participants. One Remote participant found them "clear and easily understandable," while another complained that "some instructions are very verbose and I wanted to skip reading them." IT Professionals demanded more hand-holding than others, one asking for "a demo session first to know what to expect" and another stating that "I would be okay with a little bit more information." Though effort was put into displaying game progression, it appears that more clarity is needed in that regard, as one participant admitted, "I was a little confused how many times we had to play a round." These qualitative findings align with our feedback survey results as covered earlier in this section.

User Engagement and Experience

The study participants generally found the game to be engaging. Roleplay and competition drove engagement, with one HS Student stating, "I liked the competitive aspect because it motivated me to put more effort into learning" and another suggesting: "I really enjoyed almost all aspects of the game, but I like the opportunity to become a trickster for a moment the most." Engagement with the game declined after the first game session, with one participant claiming: "the first time I played it, it felt fresh, during the second time i was getting a little bit bored." When IT Professionals were polled during the focus group, only four out of eight wanted to play a third session. This decline in engagement is also reflected in our feedback survey values, where HS Students showed the lowest willingness to retrain at 50%.

5.5.3 Summary of Participant Feedback

Overall, *Phooled!* was well-received by the participants, with approximately 90% of participants rating it as fun and educational. Collaborative participants especially praised the social aspects of the game. Learning was reported irrespective of cohort, with the majority of HS Students polled stating that most of them improved their knowledge. Furthermore, Remote participants claimed that the game "taught me more about scam emails."

A ceiling effect on learning may be present, as some subject matter experts reported that the training did not improve their phishing detection skills beyond their prior knowledge. The gamification aspects were well-liked and participants enjoyed playing against the PhishBot. Participants would, however, prefer fewer round rotations within a game session. Participants noted that the time pressure aspect forced their phishing detection decisions inside the game. The consensus split about game instructions suggested that the game instructions could be more simple or made more implicit within the game itself.

Chapter 6

Conclusions and Future Work

In this study, we investigated whether collaborative anti-phishing training through *Phooled!* improves learners' phishing detection, and examined the role of individual characteristics in predicting this improvement. We further evaluated the in-game interaction mechanics of *Phooled!* and their effectiveness. Finally, we examined which email features contribute to phishing detection difficulty.

We found that Collaborative training with *Phooled!* improved phishing detection speed and efficiency, while accuracy remained the same. Single-Player training showed no measurable improvement over the Control group. Participants in the Collaborative treatment completed their detection assessments 29% faster and 32% more efficiently than the Control group, supporting H2. H1 is therefore partially supported. The lack of accuracy improvement is likely due to the fact that most participants' baseline accuracy of 75% was on par with or slightly higher than those of prior studies (65–71%) [20, 61, 90], leaving less room for measurable gains.

We also examined whether individual characteristics predicted phishing detection improvement after training. In particular, higher levels of Motivation and Perceived Ability consistently predicted greater improvement in phishing detection efficiency, regardless of cohort or treatment condition, supporting H4. Thus, our findings support the choice of the MPA construct derived in Chapter 3 as a useful measure of motivation in the anti-phishing

training context.

Subjective norms also predicted detection improvement, but the two scales showed differing effects: peer influence was associated with worse performance, while authority influence was associated with better performance, supporting H3. Collaborative gameplay, however, eliminated the effect of both subjective norms scales, suggesting that a collaborative setting may neutralise these opposing influences on learning. Similarly, the effect of Test Anxiety varied by cohort and treatment condition: higher anxiety hindered IT Professionals' performance, while Collaborative gameplay mitigated this effect.

We evaluated the in-game interaction mechanics that comprise *Phooled!* and found that while cue detection rates improved over exposure to learning material, transfer of knowledge between players, specifically between the expert and the novice, was not visible. Furthermore, while no individual metric was linked to phishing detection improvement, in totality, they contributed to improved engagement. Our analysis of the role-play mechanism showed that passive roles performed comparably to active roles, suggesting that incorporating passive roles in the game for spectating players sustains engagement and thereby encourages continued learning.

Finally, we found that Safe emails were consistently harder to classify than Phishing emails, suggesting that email type was the primary determinant of difficulty. This suggests that future games could benefit from a higher proportion of Safe emails presented to the user, to provide learners with more exposure to ambiguous Safe emails and avoid false positives. Most participants processed the persuasive messages heuristically, relying on surface-level indicators rather than systematically evaluating the cues. Moreover, spending more time evaluating the email did not improve accuracy. Participants overestimated their own phishing detection ability, suggesting that a user's internal calibration of their skills might limit their capacity to apply them in practice.

6.1 Contributions to Anti-Phishing Education and Game Design

This thesis makes two primary design contributions to the anti-phishing education domain. First, we derived the MPA construct as a reduced subset of the MSLQ survey, based on a study conducted from a university engineering course across different instruction modes. We then applied this construct to the anti-phishing education domain and demonstrated its influence in phishing detection improvement after training.

Secondly, we designed and implemented *Phooled!*, which, as far as we are aware, is the first digital anti-phishing game with remote and collaborative play capability. We show that remote collaborative play is an effective instruction method by finding that Collaborative training improved phishing detection efficiency by 32% compared to the Control group, while Single-Player training showed no measurable improvement. We suggest that the Inverse Efficiency Score (IES) is a useful composite metric for measuring speed-accuracy tradeoffs in anti-phishing education.

The design and implementation of *Phooled!* provides several methodological contributions. We developed a semi-automated email labelling pipeline to label real-world emails to power our game and its feedback mechanisms. This pipeline enables our game to display holistic email content, covering the full email body and metadata, rather than URLs alone.

We designed a rich set of in-game interaction mechanics grounded in cognitive psychology and learning theory, including time pressure to simulate real-world email reading conditions, role-play to encourage perspective shifts, and a collaborative group cue detection mechanism to encourage knowledge transfer within the group. We show that passive spectator roles encourage user engagement levels comparable to active roles. Finally, we showcase a real-time client-server architecture capable of synchronisation of remote clients with support for concurrent game sessions, ensuring the scalability of the implementation for large organisations.

Limitations and Threats to Validity As mentioned in Section 5.2.4, differences among instruction modes may reflect cohort differences rather than delivery method. The sample size of our study is 64 participants, therefore, analysing the data across three cohorts and three treatment conditions limits the statistical power of our inference tests. The in-game interaction mechanics of *Phooled!* are inherently intertwined, which makes it difficult for us to analyse the contribution of individual mechanics. Therefore, our analysis reflects the total contribution of all mechanics.

The IT Professionals cohort completed only a single game session and therefore did not take the second Post-Test. The mixed-effects models used in our analyses are, however, robust to this imbalance and do not bias our estimates. The Control condition was also unevenly distributed across cohorts. As explained in Section 4.2, a Control group was included only in the larger Remote cohort ($n = 35$). The comparison between the Control group and the other conditions is therefore based on this cohort, and its generalisation to the other cohorts remains to be confirmed. Ideally, each cohort would contain its own Control group, and we recommend such a design for future studies.

Future Work We suggest that future studies should include more participants to improve statistical power, enabling the isolation of individual in-game interaction mechanics. To increase engagement in the Single-Player mode, PhishBot could be extended with a Large Language Model (LLM) to model an intelligent opponent rather than making pseudo random decisions. Furthermore, this LLM could be integrated into the chat to provide feedback and tips.

To measure retention of learned information, a longitudinal study over a longer time horizon could test whether improvements persist after the initial training effect. Specifically, retention could be tested using an embedded phishing attack to validate real-world users' heuristic behaviour during routine email processing. The low turnout observed when recruiting participants from the general university population suggests that future studies targeting this demographic should offer compensation or rewards to boost participation levels.

References

[1] Phishducation: Learn to outsmart email scams. https://phishducation.ca/, 2020. Accessed: 2026-05-14.

[2] Hossein Abroshan, Jan Devos, Geert Poels, and Eric Laermans. COVID-19 and Phishing: Effects of Human Emotions, Behavior, and Demographics on the Success of Phishing Attempts During the Pandemic. *IEEE Access*, 9:121916–121929, 2021.

[3] Icek Ajzen. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179–211, December 1991.

[4] Naci Akdemir and Serkan Yenal. How Phishers Exploit the Coronavirus Pandemic: A Content Analysis of COVID-19 Themed Phishing Emails. *Sage Open*, 11(3):21582440211031879, July 2021.

[5] Ali F. Al-Qahtani and Stefano Cresci. The COVID-19 scamdemic: A survey of phishing attacks and their countermeasures during COVID-19. *IET Information Security*, 16(5):324–345, 2022.

[6] Abdullah Alnajim and Malcolm Munro. An Anti-Phishing Approach that Uses Training Intervention for Phishing Websites Detection. In *2009 Sixth International Conference on Information Technology: New Generations*, pages 405–410, April 2009.

[7] Ahmed Alyahya and George R S Weir. Understanding Responses to Phishing in Saudi Arabia via the Theory of Planned Behaviour. In *2021 National Computing Colleges Conference (NCCC)*, pages 1–6, March 2021.

[8] Samer Atawneh and Hamzah Aljehani. Phishing Email Detection Model Using Deep Learning. *Electronics*, 12(20), October 2023.

[9] Aurelien Baillon, Jeroen de Bruin, Aysil Emirmahmutoglu, Evelien van de Veer, and Bram van Dijk. Informing, simulating experience, or both: A field experiment on phishing risks. *PLoS ONE*, 14(12):e0224216, December 2019.

[10] Albert Bandura and Nancy E. Adams. Analysis of self-efficacy theory of behavioral change. *Cognitive Therapy and Research*, 1(4):287–310, December 1977.

[11] Christian Bergh and Marc J. Dupuis. Cybersecurity is Stressful: The Impact of Stress on Identifying Phishing Attacks. In *2025 Cyber Awareness and Research Symposium (CARS)*, pages 1–8, Grand Forks, ND, USA, October 2025. IEEE.

[12] Elyssa Boulila, Marc Dacier, Siva Prem Vengadessa Peroumal, Nicolas Veys, and Simone Aonzo. A Closer Look At Modern Evasive Phishing Emails. In *2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 650–663, June 2025.

[13] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, January 2006.

[14] Marcus Butavicius, Ronnie Taib, and Simon J. Han. Why people keep falling for phishing scams: The effects of time pressure and deception cues on the detection of phishing emails. *Computers & Security*, 123:102937, December 2022.

[15] Marcus A Butavicius, Kathryn Parsons, Malcolm R Pattinson, Agata McCormac, Dragana Calic, and Meredith Lillie. Understanding susceptibility to phishing emails: Assessing the impact of individual differences and culture. *HAISA*, 17:12–23, 2017.

[16] Casey Inez Canfield, Baruch Fischhoff, and Alex Davis. Quantifying Phishing Susceptibility for Detection and Behavior Decisions. *Human Factors*, 58(8):1158–1172, December 2016.

[17] Shelly Chaiken. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology*, 39(5):752, 1980.

[18] Catherine Chen and Joel Whitesel. The Validity and Reliability Study of a Revised Motivated Strategy for Learning Questionnaire (MSLQ) for Assessing Computer Software Learning Strategies. *International Journal of E-Adoption (IJEA)*, 4(2):28–51, April 2012.

[19] Xiaowei Chen, Margault Sacre, Gabriele Lenzini, Samuel Greiff, Verena Distler, and Anastasia Sergeeva. The effects of group discussion and role-playing training on self-efficacy, support-seeking, and reporting phishing emails: Evidence from a mixed-design experiment. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024.

[20] Gokul CJ, Sankalp Pandit, Sukanya Vaddepalli, Harshal Tupsamudre, Vijayanand Banahatti, and Sachin Lodha. PHISHY - A Serious Game to Train Enterprise Users on Phishing Awareness. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, CHI PLAY '18 Extended Abstracts, pages 169–181, New York, NY, USA, October 2018. Association for Computing Machinery.

[21] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, New York, 2 edition, May 2013.

[22] David A Cook, Warren G Thompson, and Kris G Thomas. The Motivated Strategies for Learning Questionnaire: score validity among medicine residents. *Medical education*, 45(12):1230–1240, 2011.

[23] Christine P. Dancey and John Reidy. *Statistics without maths for psychology*. Pearson, Prentice Hall, Harlow, 5. ed edition, 2011.

[24] David Knox and Vignesh Kumar Karuppasamy. Motivation and Learning Strategies - What can university engineering design courses do to help students and what must

students do? In *2025 ASEE Annual Conference & Exposition*, number 10.18260/1-2–56977, Montreal, Quebec, Canada, June 2025. ASEE Conferences.

[25] Edward L Deci and Richard M Ryan. The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry*, 11(4):227–268, 2000.

[26] 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*, eCrime '07, pages 37–44, New York, NY, USA, October 2007. Association for Computing Machinery.

[27] 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, pages 79–90, New York, NY, USA, July 2006. Association for Computing Machinery.

[28] Angela L. Duckworth, Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92(6):1087–1101, 2007.

[29] W. Holmes Finch. Introduction to Factor Analysis. In *Exploratory Factor Analysis*, pages 1–12. SAGE Publications, Inc., 2020.

[30] BJ Fogg. A behavior model for persuasive design. In *Proceedings of the 4th International Conference on Persuasive Technology*, Persuasive '09, pages 1–7, New York, NY, USA, April 2009. Association for Computing Machinery.

[31] Giorgio Fumera, Ignazio Pillai, and Fabio Roli. Spam filtering based on the analysis of text information embedded into images. *Journal of Machine Learning Research*, 7(12), 2006.

[32] Brandon E. Gavett, Rui Zhao, Samantha E. John, Cara A. Bussell, Jennifer R. Roberts, and Chuan Yue. Phishing suspiciousness in older and younger adults: The role of executive functioning. *PLOS ONE*, 12(2):e0171620, February 2017.

[33] Calixto Gutiérrez-Braojos, Jesus Montejo-Gamez, Ana Marin-Jimenez, and Jesús Campaña. Hybrid learning environment: Collaborative or competitive learning? *Virtual Reality*, 23(4):411–423, December 2019.

[34] Lukas Hafner, Florian Wutz, Daniela Pohn, and Wolfgang Hommel. TASEP: A Collaborative Social Engineering Tabletop Role-Playing Game to Prevent Successful Social Engineering Attacks. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, pages 1–10, Benevento Italy, August 2023. ACM.

[35] Tzipora Halevi, Nasir Memon, and Oded Nov. Spear-Phishing in the Wild: A Real-World Study of Personality, Phishing Self-Efficacy and Vulnerability to Spear-Phishing Attacks. *SSRN Electronic Journal*, 2015.

[36] Florian Hartig. DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models, August 2016.

[37] Lauren C Hensley, Ryan Iaconelli, and Christopher A Wolters. "This weird time we're in": How a sudden change to remote education impacted college students' self-regulated learning. *Journal of Research on Technology in Education*, 54(sup1):S203–S218, 2022.

[38] Jan Herrington, Thomas C. Reeves, and Ron Oliver. *Authentic Learning Environments*, pages 401–412. Springer, New York, NY, 2014.

[39] Jonathan C. Hilpert, Jennifer Stempien, Katrien J. van der Hoeven Kraft, and Jenefer Husman. Evidence for the Latent Factor Structure of the MSLQ: A New Conceptualization of an Established Questionnaire. *SAGE Open*, 3(4):215824401351030, January 2013.

[40] Grant Ho, Ariana Mirian, Elisa Luo, Khang Tong, Euyhyun Lee, Lin Liu, Christopher A. Longhurst, Christian Dameff, Stefan Savage, and Geoffrey M. Voelker. Understanding the efficacy of phishing training in practice. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 37–54, San Francisco, CA, USA, May 2025. IEEE.

[41] Alastair Irons and Sam Elkington. *Enhancing Learning through Formative Assessment and Feedback*. Routledge, London, 2 edition, August 2021.

[42] Mohammad S. Jalali, Maike Bruckes, Daniel Westmattelmann, and Gerhard Schewe. Why Employees (Still) Click on Phishing Links: Investigation in Hospitals. *Journal of Medical Internet Research*, 22(1):e16775, January 2020.

[43] Daniel Jampen, Gurkan Gur, Thomas Sutter, and Bernhard Tellenbach. Don't click: towards an effective anti-phishing training. A comparative literature review. *Human-centric Computing and Information Sciences*, 10(1):33, August 2020.

[44] Gokul Jayakrishnan, Vijayanand Banahatti, and Sachin Lodha. PickMail: A Serious Game for Email Phishing Awareness Training. In *Proceedings 2022 Symposium on Usable Security*, San Diego, CA, 2022. Internet Society.

[45] Matthew L. Jensen, Michael Dinger, Ryan T. Wright, and Jason Bennett Thatcher. Training to Mitigate Phishing Attacks Using Mindfulness Techniques. *Journal of Management Information Systems*, 34(2):597–626, April 2017.

[46] David W Johnson, Roger T Johnson, and Mary Beth Stanne. Cooperative Learning Methods: A Meta-Analysis. 2000.

[47] Joseph F. Hair Jr, William C. Black, Barry J. Babin, and Rolph E. Anderson. *Multivariate Data Analysis*. Pearson, Upper Saddle River, NJ, 2010.

[48] Menelaos N. Katsantonis, Athanasios Manikas, Nikolaos Partarakis, and Ioannis Mavridis. Yellow training: a collaborative tCOFELET tabletop serious game for cybersecurity awareness and training. *International Journal of Information Security*, 24(4):171, July 2025.

[49] Kristian Kiili. Digital game-based learning: Towards an experiential gaming model. *The Internet and Higher Education*, 8(1):13–24, January 2005.

[50] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. School of phish: a real-world

evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, pages 1–12, New York, NY, USA, July 2009. Association for Computing Machinery.

[51] Ponnurangam Kumaraguru, Yong Rhee, Steve Sheng, Sharique Hasan, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Getting users to pay attention to anti-phishing education: evaluation of retention and transfer. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, eCrime '07, pages 70–81, New York, NY, USA, October 2007. Association for Computing Machinery.

[52] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Lessons from a real world evaluation of anti-phishing training. In *2008 eCrime Researchers Summit*, pages 1–12, Atlanta, GA, USA, October 2008. IEEE.

[53] Charlie Marriott. Through the Net: Investigating How User Characteristics Influence Susceptibility to Phishing. *Dissertations*, January 2018.

[54] Lindsay McCardle and Allyson F Hadwin. Using multiple, contextualized data sources to measure learners' perceptions of their self-regulated learning. *Metacognition and Learning*, 10(1):43–75, 2015.

[55] Steven McElwee, George Murphy, and Paul Shelton. Influencing Outcomes and Behaviors in Simulated Phishing Exercises. In *SoutheastCon 2018*, pages 1–6, April 2018.

[56] Barbara Means, Yukie Toyama, Robert Murphy, and Marianne Baki. The Effectiveness of Online and Blended Learning: A Meta-Analysis of the Empirical Literature. *Teachers College Record: The Voice of Scholarship in Education*, 115(3):1–47, March 2013.

[57] Gaurav Misra, Nalin Asanka Gamagedara Arachchilage, and Shlomo Berkovsky. Phish Phinder: A Game Design Approach to Enhance User Confidence in Mitigating Phishing Attacks. *arXiv:1710.06064 [cs]*, October 2017.

[58] Gregory D. Moody, Dennis F. Galletta, and Brian Kimball Dunn. Which phish get caught? An exploratory study of individuals' susceptibility to phishing. *European Journal of Information Systems*, 26(6):564–584, November 2017.

[59] 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, May 2019.

[60] Emily Naul and Min Liu. Why story matters: A review of narrative in serious games. *Journal of Educational Computing Research*, 58(3):687–707, 2020.

[61] Cas Pars. PHREE of Phish : The Effect of Anti-Phishing Training on the Ability of Users to Identify Phishing Emails, July 2017.

[62] Kathryn Parsons, Marcus Butavicius, Paul Delfabbro, and Meredith Lillie. Predicting susceptibility to social influence in phishing emails. *International Journal of Human-Computer Studies*, 128:17–26, August 2019.

[63] Gordon Pennycook. A perspective on the theoretical foundation of dual process models. In *Dual process theory 2.0.*, Current issues in thinking and reasoning, pages 5–27. Routledge/Taylor & Francis Group, New York, NY, US, 2018.

[64] PR Pintrich. A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ). *National Center for Research to Improve Postsecondary Teaching and Learning*, 1991.

[65] Proofpoint. Cost of Phishing Scams Triples Since 2015. https://www.proofpoint.com/us/blog/security-awareness-training/new-ponemon-study-finds-annual-cost-phishing-scams-has-more-tripled, 2021. Accessed: 2026-05-14.

[66] Proofpoint. 2024 State of the Phish. https://www.proofpoint.com/uk/resources/threat-reports/state-of-phish, 2024. Accessed: 2026-05-14.

[67] Jhon Jairo Ramirez Echeverry, Agueda Garcia Carrillo, and Fredy Andres Olarte Dussan. Adaptation and validation of the motivated strategies for learning questionnaire-mslq-in engineering students in Colombia. *International journal of engineering education*, 32(4):1774–1787, August 2016.

[68] Rene Roepke, Vincent Drury, Ulrike Meyer, and Ulrik Schroeder. Exploring and Evaluating Different Game Mechanics for Anti-Phishing Learning Games. *International Journal of Serious Games*, 9(3):23–41, September 2022.

[69] Rene Roepke, Klemens Koehler, Vincent Drury, Ulrik Schroeder, Martin R. Wolf, and Ulrike Meyer. A Pond Full of Phishing Games - Analysis of Learning Games for Anti-Phishing Education. In George Hatzivasilis and Sotiris Ioannidis, editors, *Model-driven Simulation and Training Environments for Cybersecurity*, Lecture Notes in Computer Science, pages 41–60, Cham, 2020. Springer International Publishing.

[70] Dawn M. Sarno and Mark B. Neider. So many phish, so little time: Exploring email task factors and phishing susceptibility. *Human Factors*, 64(8):1379–1403, December 2022.

[71] Hamidreza Shahbaznezhad, Farzan Kolini, and Mona Rashidirad. Employees' Behavior in Phishing Attacks: What Individual, Organizational, and Technological Factors Matter? *Journal of Computer Information Systems*, pages 1–12, October 2020.

[72] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith 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*, SOUPS '07, pages 88–99, New York, NY, USA, July 2007. Association for Computing Machinery.

[73] Miguel Sicart. Defining game mechanics. *Game studies*, 8(2):1–14, 2008.

[74] Michelle Steves, Kristen Greene, and Mary Theofanos. Categorizing human phishing difficulty: a Phish Scale. *Journal of Cybersecurity*, 6(1):tyaa009, January 2020.

[75] Jerry Chih-Yuan Sun, Cian-Yu Kuo, Huei-Tse Hou, and Yu-Yan Lin. Exploring Learners' Sequential Behavioral Patterns, Flow Experience, and Learning Performance in an Anti-Phishing Educational Game. *Educational Technology & Society*, 20(1):45–60, 2017.

[76] Jerry Chih-Yuan Sun and Hsiu-Shan Lin. Effects of integrating an interactive response system into flipped classroom instruction on students' anti-phishing self-efficacy, collective efficacy, and sequential behavioral patterns. *Computers & Education*, 180:104430, April 2022.

[77] Barbara G Tabachnick, Linda S Fidell, and Jodie B Ullman. *Using multivariate statistics*, volume 5. pearson Boston, MA, 2007.

[78] Mohsen Tavakol and Reg Dennick. Making sense of Cronbach's alpha. *International journal of medical education*, 2:53, 2011.

[79] Franklin Tchakounte, Leonel Kanmogne Wabo, and Marcellin Atemkeng. A Review of Gamification Applied to Phishing, March 2020.

[80] Kutub Thakur, Md Liakat Ali, Muath A. Obaidat, and Abu Kamruzzaman. A Systematic Review on Deep-Learning-Based Phishing Email Detection. *Electronics*, 12(21), November 2023.

[81] Valerie A. Thompson. Why It Matters: The Implications of Autonomous Processes for Dual Process Theories—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, 8(3):253–256, May 2013.

[82] James T Townsend and F Gregory Ashby. *Stochastic modeling of elementary psychological processes*. CUP Archive, 1983.

[83] James T Townsend and F Gregory Ashby. Methods of modeling capacity in simple processing systems. In *Cognitive theory*, pages 199–239. Psychology Press, 2014.

[84] Selen Turkay, Daniel Hoffman, Charles K. Kinzer, Pantiphar Chantes, and Christopher Vicari. Toward Understanding the Potential of Games for Learning: Learning Theory,

Game Design Characteristics, and Situating Video Games in Classrooms. *Computers in the Schools*, 31(1-2):2–22, April 2014.

[85] Sylke Vandercruysse, Mieke Vandewaetere, Frederik Cornillie, and Geraldine Clarebout. Competition and students' perceptions in a game-based language learning environment. *Educational Technology Research and Development*, 2013.

[86] Arun Vishwanath, Brynne Harrison, and Yu Jie Ng. Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility. *Communication Research*, 45(8):1146–1166, December 2018.

[87] 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, June 2011.

[88] Mark Watson, Michelle McSorley, Cheryl Foxcroft, and Andrea Watson. Exploring the motivation orientation and learning strategies of first year university learners. *Tertiary Education and management*, 10(3):193–207, 2004.

[89] Zikai Alex Wen, Yiming Li, Reid Wade, Jeffrey Huang, and Amy Wang. What.Hack: Learn Phishing Email Defence the Fun Way. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 234–237, Denver Colorado USA, May 2017. ACM.

[90] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. What. hack: engaging anti-phishing training through a role-playing phishing simulation game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[91] Kim Witte. Putting the fear back into fear appeals: The extended parallel process model. *Communication Monographs*, 59(4):329–349, December 1992.

[92] Olga A. Zielinska, Rucha Tembe, Kyung Wha Hong, Xi Ge, Emerson Murphy-Hill, and Christopher B. Mayhorn. One Phish, Two Phish, How to Avoid the Internet Phish: Analysis of Training Strategies to Detect Phishing Emails, 2014.

APPENDICES

Appendix A

Survey Instruments

A.1 Calibration Survey

Items marked with * had their question wording adapted for the anti-phishing game context and for cohort differences. These differences are specified in brackets for applicable questions.

Heuristic Processing

3 items, 5-point scale (Strongly Agree to Strongly Disagree). Adapted from the SCAM [86].

1. I skim through my emails.
2. I briefly look at the sender/source of the email.
3. I ignore the email message content.

Systematic Processing

3 items, 5-point scale (Strongly Agree to Strongly Disagree). Adapted from the SCAM [86].

1. I think about the action I will take based on what I see in the email.
2. I find myself making connections between the email's request and what I have heard about emails requesting such information.
3. I spend some time thinking about the request in the email before I make my decision of what action to take.

Subjective Norms 1

3 items, 5-point scale (Strongly Agree to Strongly Disagree). From Jalali et al. [42].

1. People who influence my behaviour would think that I should follow the [organisation/school]'s email security policy.*
2. People whose opinions are important to me would think that I should follow the [organisation/school]'s email security policy.*
3. People whom I respect would think that I should follow the [organisation/school]'s email security policy.*

Subjective Norms 2

4 items, 7-point scale (Strongly Agree to Strongly Disagree). From Shahbaznezhad et al. [71].

1. My [instructor/boss/supervisor] thinks that I should follow the [organisation/school]'s email security policy.*
2. My [peers/colleagues] think that I should follow the [organisation/school]'s email security policy.*
3. My [organisation/school]'s IT department pressures me to follow the [organisation/school]'s email security policy.*

4. My [juniors or underclassmen/subordinates] think I should follow the [organisation/school]'s email security policy.*

Motivation and Perceived Ability

26 items across five subscales, 7-point scale (Not at all true to Very true). Adapted from the MSLQ [64].

Intrinsic Goal Orientation (4 items).

1. In a phishing educational game I prefer to learn about phishing avoidance techniques that really challenge me so I can detect current and future types of attacks.
2. In a phishing educational game, I prefer to learn about phishing avoidance techniques that arouse my curiosity even if it is difficult to play/learn.
3. The most satisfying thing for me in a phishing educational game is to learn how to avoid all types of phishing emails, including the trickiest ones.
4. When I have the opportunity in a phishing educational game, I want to interact with tricky or difficult phishing emails that I can learn from even if I don't get a "good" score.

Extrinsic Goal Orientation (4 items).

1. Getting a "good" score is the most satisfying thing for me in a phishing educational game.
2. The most important thing for me right now is increasing my overall performance on the Phishducation platform (e.g. gold coins from the game, scores on the quiz, etc.)
3. If I can, I want to get better scores than other participants in this phishing educational game.

4. I want to do well in this phishing educational game because it is important to show my ability to my family, friends, employer, or others.

Task Value (6 items).

1. I think I will be able to use what I learn from this phishing educational game in other environments besides [school/workplace].*
2. It is important for me to learn how to avoid phishing emails through this phishing educational game.
3. I am very interested in learning how to distinguish between phishing and non-phishing emails through this phishing educational game.
4. I think the content of this phishing educational game would be useful for me to learn.
5. I like learning phishing avoidance techniques.
6. Understanding phishing avoidance techniques is very important to me.

Control of Learning Beliefs (4 items).

1. If I practise in appropriate ways, then I will be able to get better at phishing email avoidance.
2. It is my own fault if I don't learn how to avoid phishing emails using this phishing educational game.
3. If I try hard enough, then I will understand how to avoid phishing emails with this phishing educational game.
4. If I don't understand how to avoid phishing emails using this phishing educational game, it is because I didn't try hard enough.

Self-Efficacy for Learning and Performance (8 items).

1. I believe I will receive excellent scores in this phishing educational game.
2. I'm certain I can understand even the most difficult content/concepts about phishing avoidance techniques presented in this phishing educational game.
3. I'm confident I can understand the basic content/concepts about phishing avoidance techniques presented in this phishing educational game.
4. I'm confident I can understand the most complex content/concepts about phishing avoidance techniques presented in this phishing educational game.
5. I'm confident I can do an excellent job on each part/section of this phishing educational game.
6. I expect to do well in this phishing educational game.
7. I'm certain I can master the phishing avoidance techniques taught in this phishing educational game.
8. Considering the difficulty of learning phishing avoidance techniques, I think I will do well in this phishing educational game.

Test Anxiety

5 items, 7-point scale (Not at all true to Very true). Adapted from the MSLQ [64].

1. When I play a phishing educational game, I think about how poorly I am doing compared with other participants.
2. When I play a (phishing or any type of) educational game, I think about other sections/parts of the game that I'm not good at.

3. When I play a (phishing or any type of) educational game, I think of the consequences of performing badly.

4. I have an uneasy, upset feeling when I play a (phishing or any type of) educational game.

5. I feel my heart beating fast when I play a (phishing or any type of) educational game.

A.2 Post-Study Survey

The post-study feedback survey contains both quantitative Likert-based items, as well as qualitative open-ended questions. The Likert scale ranged on a 5-point scale (Strongly Agree to Strongly Disagree).

Likert Items

1. I found the *Phooled!* game to be fun/enjoyable.

2. I think the *Phooled!* game is a good educational tool to help me learn how to avoid phishing emails.

3. I would be willing to play this game as a routine educational tool a few times a year to learn the latest phishing methods used by adaptive malicious attackers.

4. I found the game instructions to be understandable.

5. I found the game interface to be usable/intuitive.

6. I prefer competitive gameplay rather than inter-player collaboration.

Open-Ended Questions

7. Which aspects of the *Phooled!* game did you like?
8. Which aspects of the *Phooled!* game did you dislike?
9. What features or changes could be implemented to enhance/improve the *Phooled!* game?
10. What is your preferred duration of a single game session of *Phooled!*? (in minutes)
11. In a group setting, what would be your preferred number of players for this game?
12. Would you enjoy playing this game in a fully online setting with randomly matched users? If no, why not?

A.3 Focus Group Discussion Guide

This semi-structured discussion guide was used to guide the in-person focus group sessions.

Phooled! Game Experience

1. What is your experience using our educational quiz/game?
2. What were the highlights of your experience using the educational quiz/game?
3. What would you change about the educational quiz/game?

Prior Knowledge

4. What is your background like? Are you tech-savvy or have a considerable online presence?

5. Tell me more about your usage of emails. Do you read or send emails a lot? How many hours per week would you say you interact with emails?
6. How "capable" of email technology do you think you are? Do you use multiple email platforms? What kind of activities do you use email for?
7. Are there any particular email habits you follow? Any rule-of-thumbs?
8. Now, tell me what you know about phishing email attacks.
9. Did you or anyone you know get targeted by a phishing attempt?
10. How did you feel during and after the phishing attack?
11. Do you feel a personal connection to other phishing victims after this attack?
12. What kind of feelings do you have toward the malicious actor who phished you?

Phishing Susceptibility and Learning

13. What is your general thought-process while going through the malicious emails?
14. Do you think any of the email habits you mentioned previously contributed to your victimisation or ability to avoid the phishing attempt?
15. What do you think you could have done to avoid such attacks in the future?
16. Would you say that you were uniquely susceptible to this attack? If so, what personal characteristics do you think make you susceptible?
17. Do you think you learned anything from this attack?
18. What would you do differently in the future to prevent such attacks?
19. Are you motivated to change your email habits to avoid future attempts?

Subjective Norms and Self-Efficacy

20. Do you think your general perception of whether others in your organisation are following good email habits, influences your decision to follow similar good habits?

21. If you think you don't need to do anything different, why is this the case?

Anti-phishing Training

22. How knowledgeable are you about educational software that help learners avoid phishing attacks? Have you ever used such software?

23. How comfortable would you be in using these educational software platforms?

24. Do you feel that such software is effective in their current form? Is there anything that you would like to see changed in these platforms?

25. What kind of phishing education do you think would be most effective for you?

Appendix B

Client-Server Messaging Protocol

The messaging protocol described in Section 4.3 uses JSON-encoded messages over WS connections. Each message is made up of a message code, represented as an integer, and a payload. The message code identifies the action to be performed, while the payload determines the context required to perform that action. This payload could be, for example, HTML email contents to display on the user's screen or a user's chat message text. Table B.1 shows the list of message codes applicable for different connection types.

Table B.1: Specification of Messages in *Phooled!*, with their codes and descriptions

Direction	**Message Code**	**Description**
Client to Server	`JOIN`	Establish connection with server
	`READY`	Signal readiness for next state
	`ACK`	Generic Acknowledge Message
	`CLASSIFY_CARD`	Classify current card as Phishing or Safe
	`SELECT_CARD`	Keep or discard current card
	`PLAY_CARD`	Select card to play as Trickster
	`SELECT_ATTR`	Select or deselect a cue

Continued on next page

Direction	Message Code	Description
	REQUEST_CARDS	Request additional cards in Quick Pick
	CARD_OPEN	Send card view telemetry
	CHAT_MESSAGE	Send a message in the chat
	FEEDBACK	Submit dispute for a card
	SELECT_AVATAR	Select player avatar
	HELP_MODAL_OPEN	Send help modal open telemetry
	HELP_MODAL_CLOSE	Send help modal close telemetry
	PAUSE_GAME	Request game pause
	RESUME_GAME	Request game resume
	ABANDON	Signal game abandonment
Host to Server	START	Request game start
	SET_SETTINGS	Configure game settings
	SET_ALIAS	Configure game display name
	SETTINGS_MODAL_OPEN	Send settings modal open telemetry
Server to Client	CARD_CONTENT	Deliver email card payload
	FLUSH_CARDS	Clear all cards in player's hand
	REQUEST_ACK	Acknowledge Generic Request
	SET_ROLE	Assign game role
	FR_ROUND_NUMBER	Signal current rotation progress
	SEND_RESULT_CLASSIFY	Provide classification ground truth
	SEND_RESULT_ATTR	Provide cue identification ground truth
	RESULT_MESSAGE	Provide feedback message
	GOLD_COINS_TEXT	Announce Gold Coins earned
	SEND_PLAYER_STATUS	Update game players' status
	SYNC_STATE	Synchronise client state
	SYNC_TIMER	Synchronise client state timer
	SYNC_SETTINGS	Broadcast updated game settings

Continued on next page

Direction	Message Code	Description
	USER_JOIN	Announce players joining game
	PLAYERS_LIST	Update connected players' statuses
	AVATARS_LIST	Update players' avatars
	SET_ALIAS	Propagate new game display name
	CHAT_MESSAGE	Broadcast a chat message
	SELECT_ATTR	Broadcast cue selection to all players
	GOOGLE_FORM_DETAILS	Redirect to post-game survey or assessment
	SHOW_ERROR	Display custom error message

Appendix C

Normality Tests

In this appendix, we present Q–Q plots used in tests of normality for the data presented in Chapter 5. The plots are broken down by cohort and treatment condition where appropriate.

C.1 Total Classifications

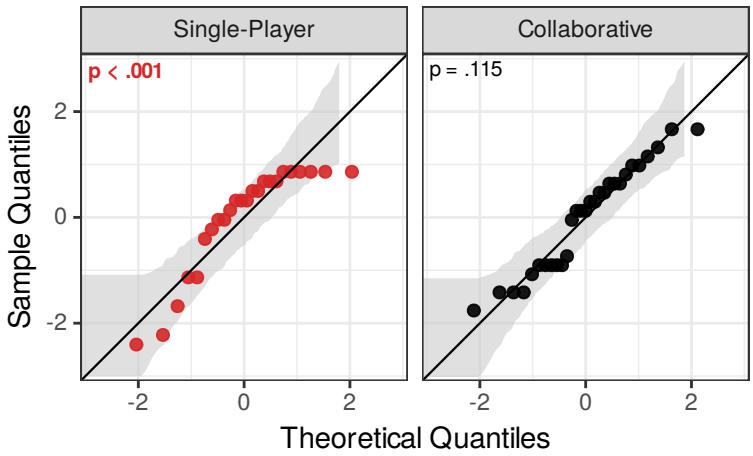


Figure C.1: Q–Q plots for total classifications by treatment condition

C.2 Game Duration

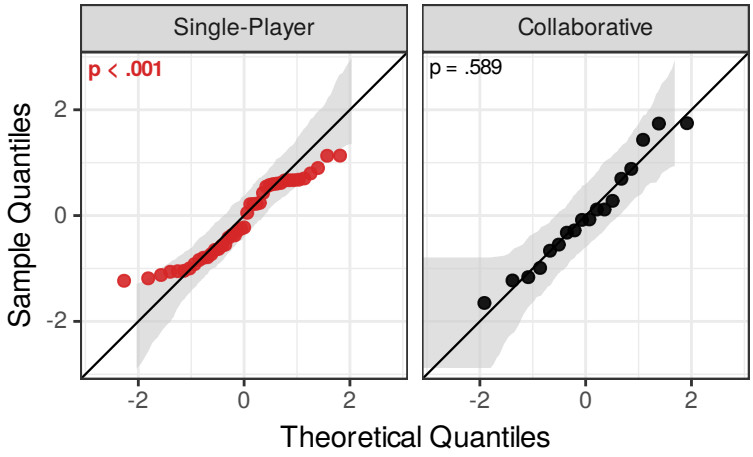


Figure C.2: Q–Q plots for game duration by treatment condition

C.3 Motivation and Perceived Ability (MPA)

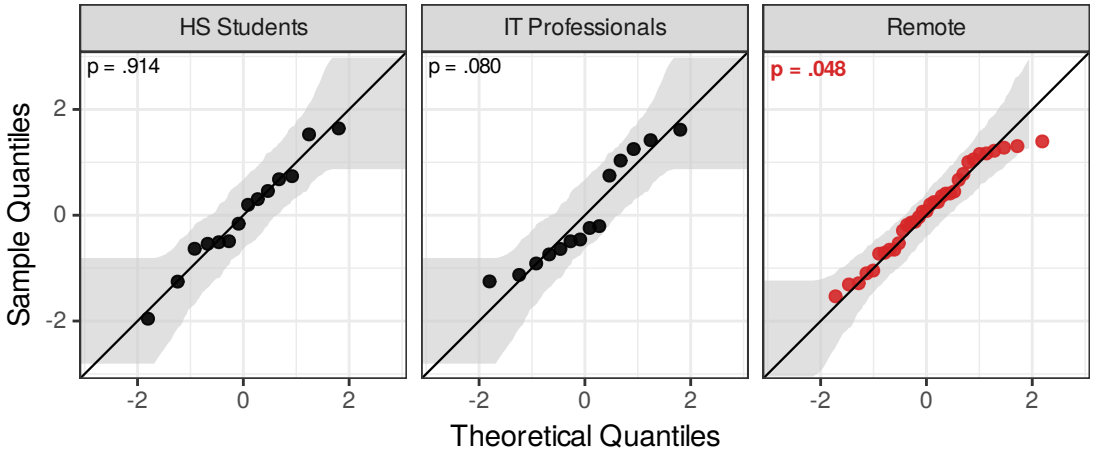


Figure C.3: Q–Q plots for MPA values by cohort

C.4 Test Anxiety

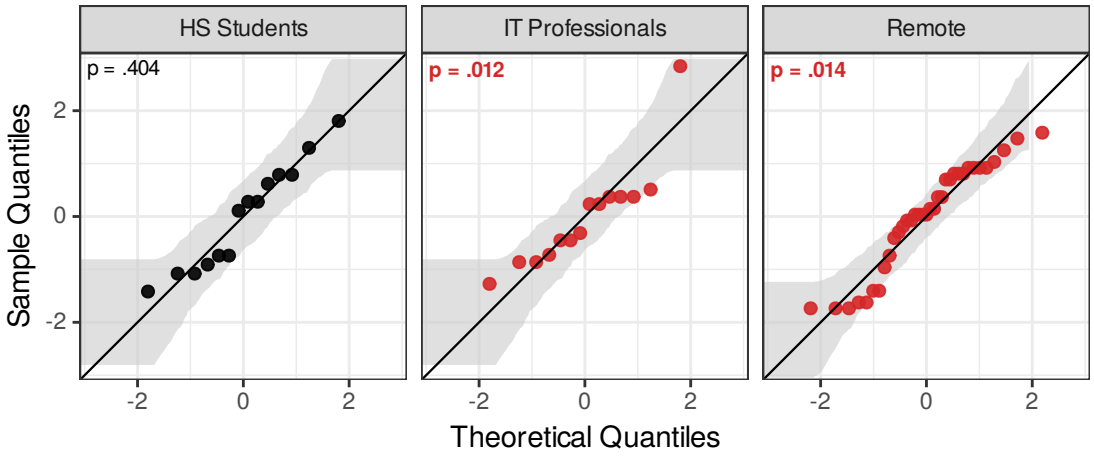


Figure C.4: Q–Q plots for Test Anxiety values by cohort

C.5 Email Sample Metrics

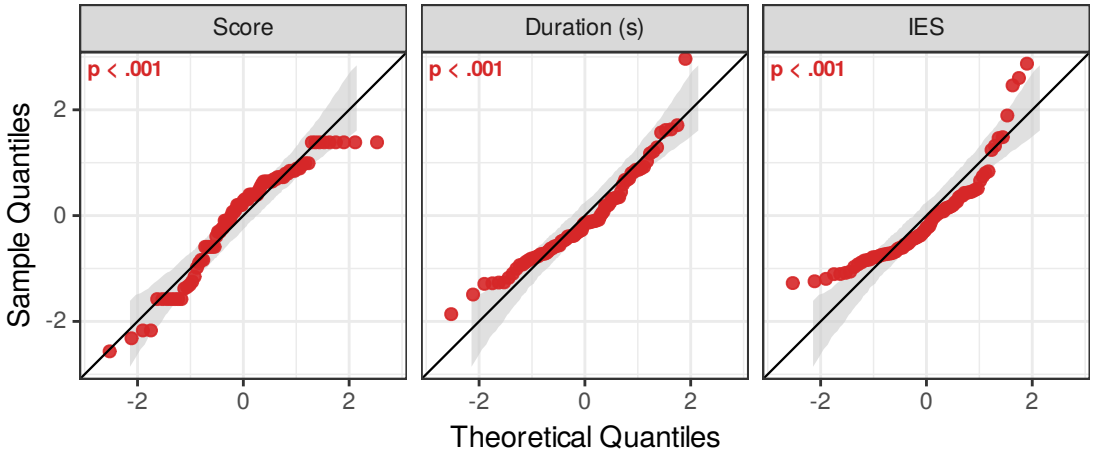


Figure C.5: Q–Q plots for email sample metrics

C.6 IES by Sample Type

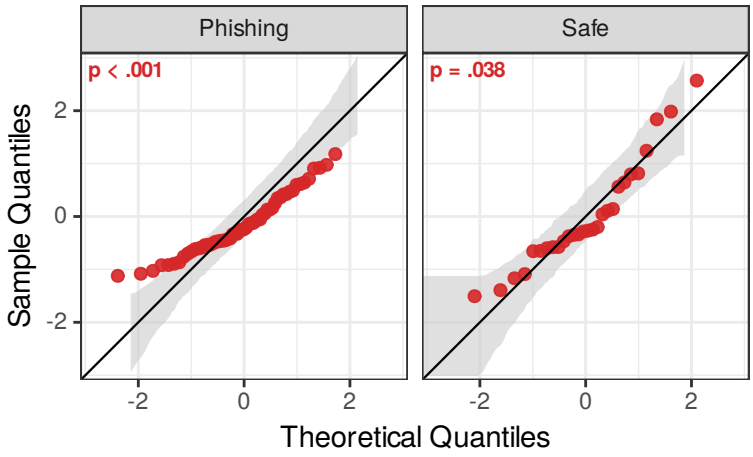


Figure C.6: Q–Q plots for IES by sample type

C.7 Card Selection and Play Ratios

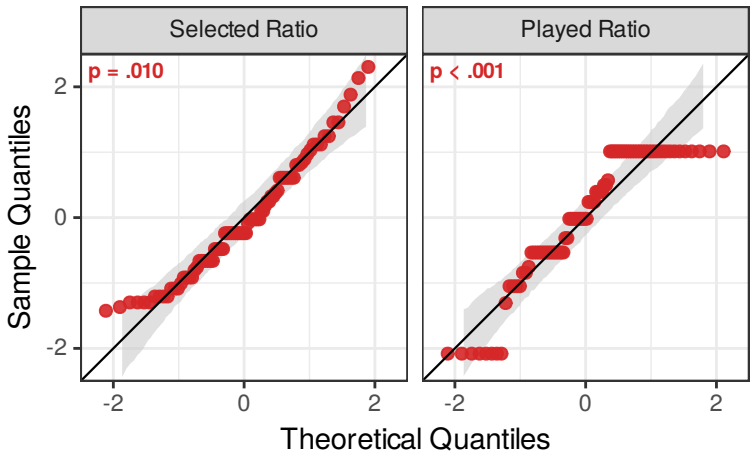


Figure C.7: Q–Q plots for card selected ratio and played ratio

C.8 Classification Duration by Game State

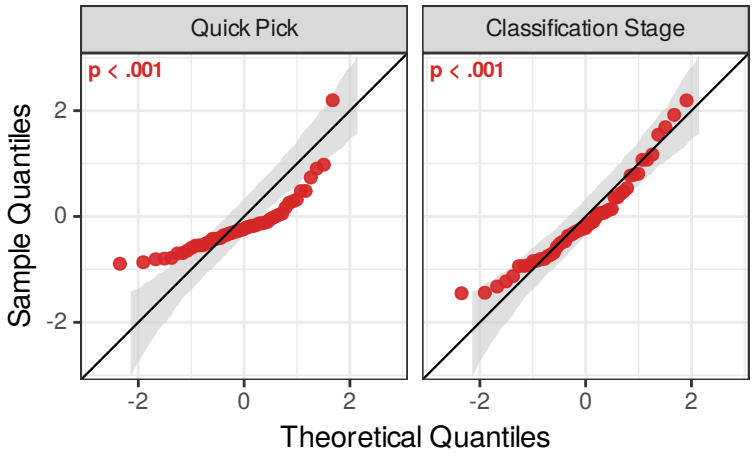


Figure C.8: Q–Q plots for classification duration by game state

C.9 IES by Role

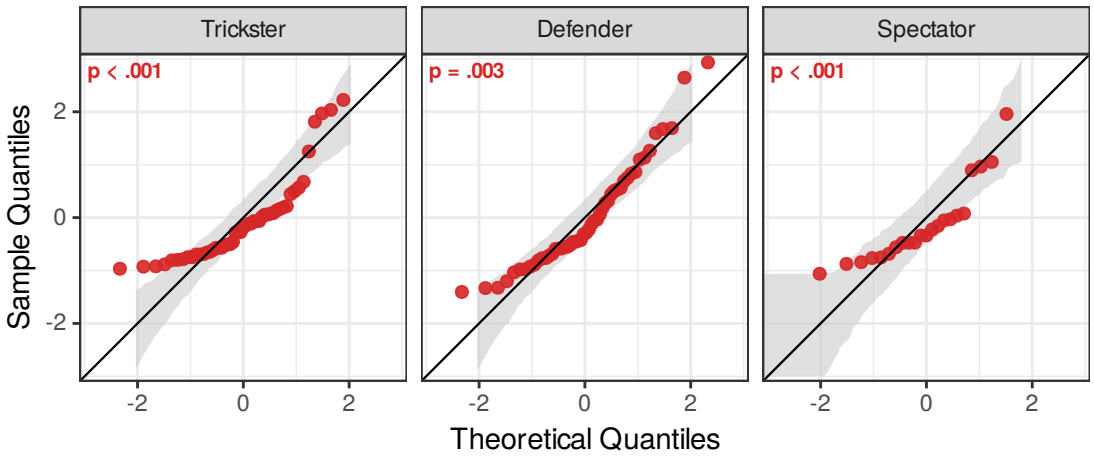


Figure C.9: Q–Q plots for IES by role

C.10 Cue F1 Scores

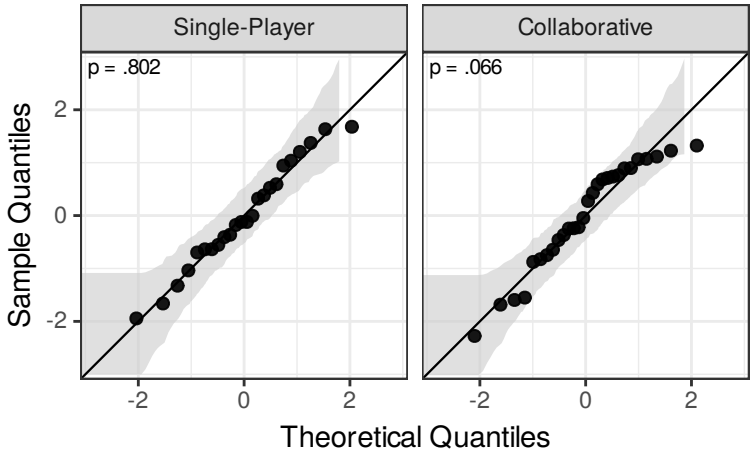


Figure C.10: Q–Q plots for cue F1 scores by treatment

C.11 Heuristic and Systematic Scales

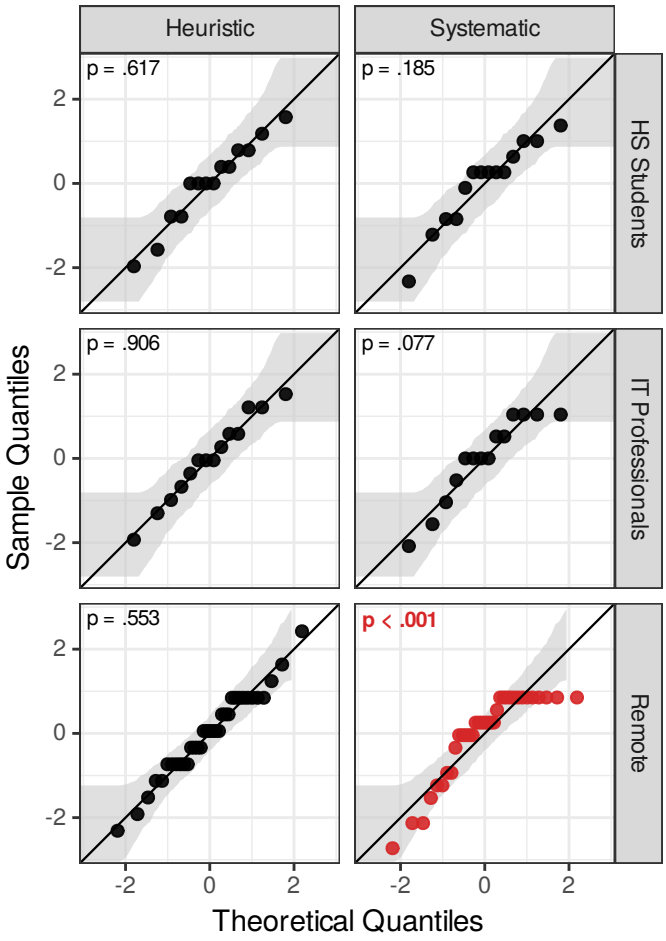


Figure C.11: Q–Q plots for Heuristic and Systematic scale values by cohort and treatment condition

C.12 Subjective Norms

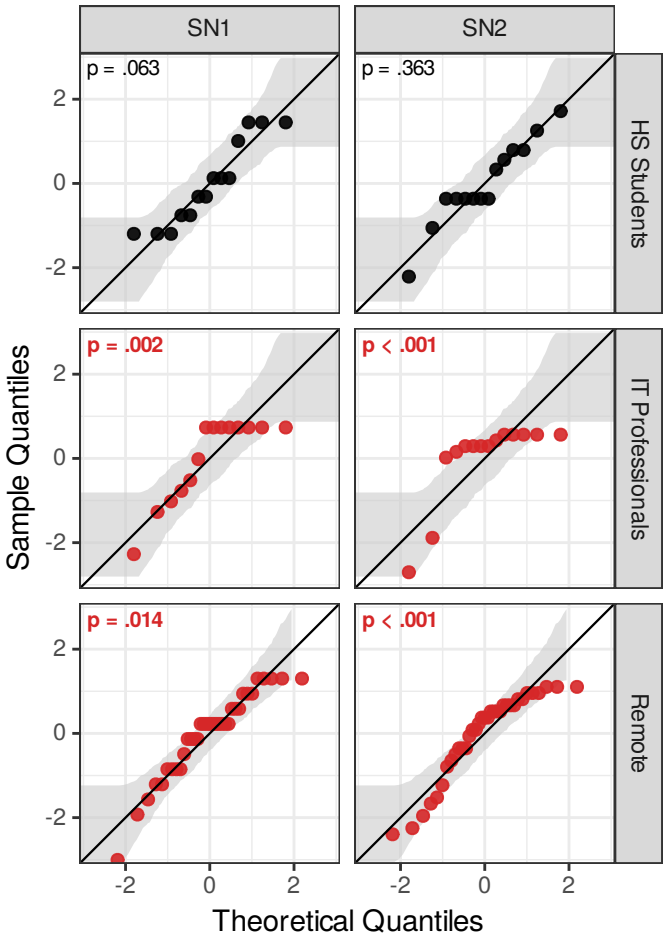


Figure C.12: Q–Q plots for subjective norms scale values by cohort and treatment condition

C.13 Phishing Detection Accuracy

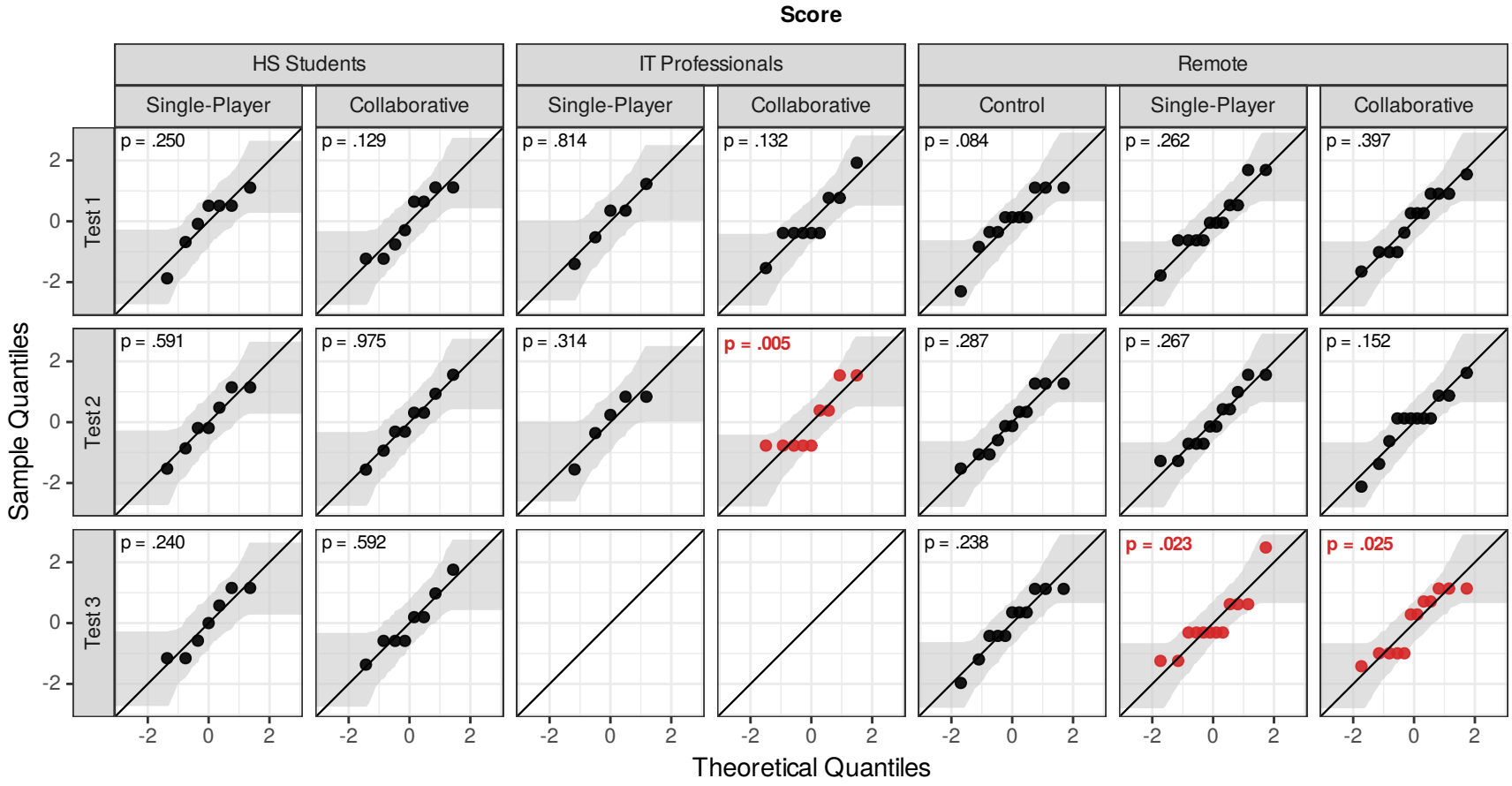


Figure C.13: Q–Q plots for test scores by cohort and treatment condition

C.14 Phishing Detection Speed

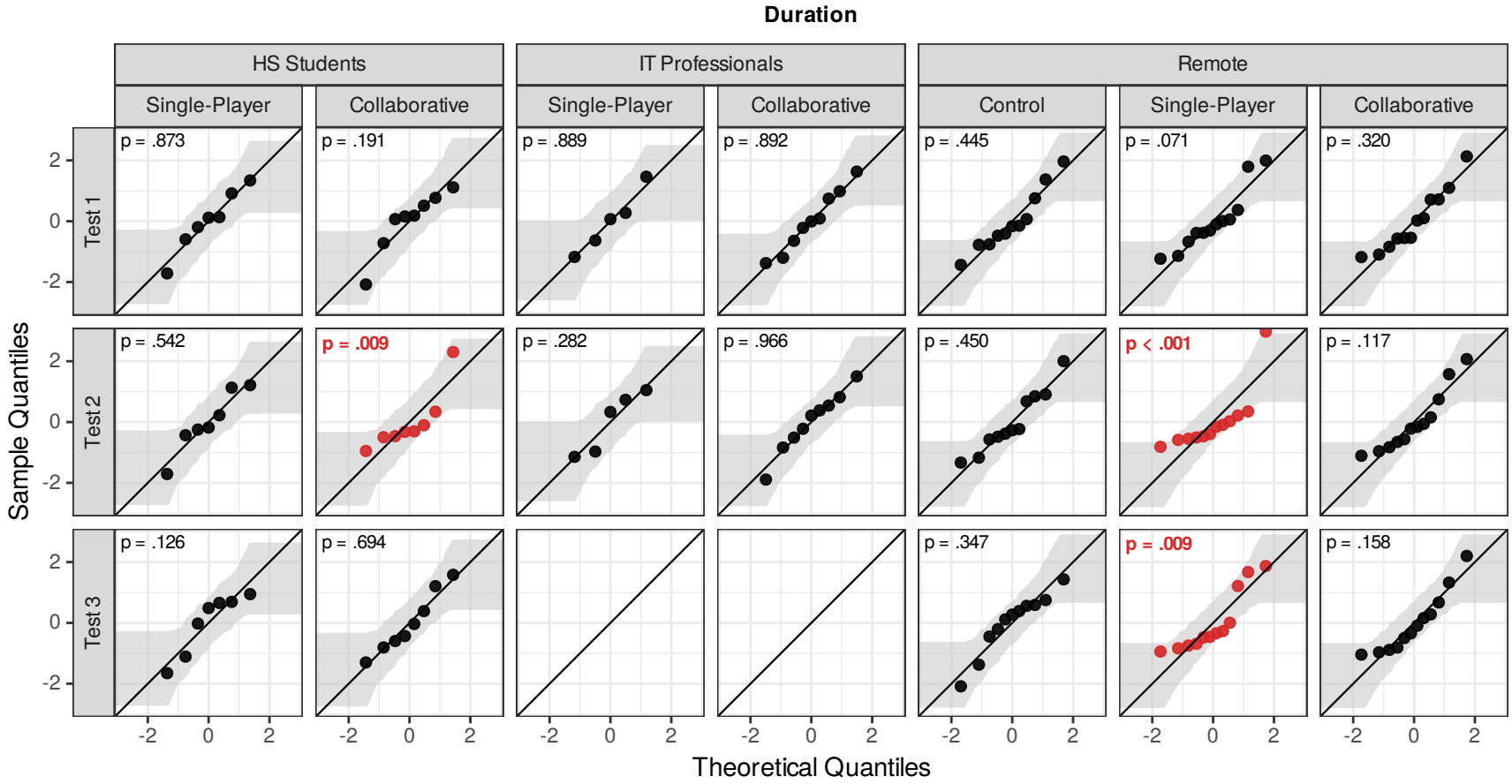


Figure C.14: Q–Q plots for test duration by cohort and treatment condition

C.15 Phishing Detection Efficiency

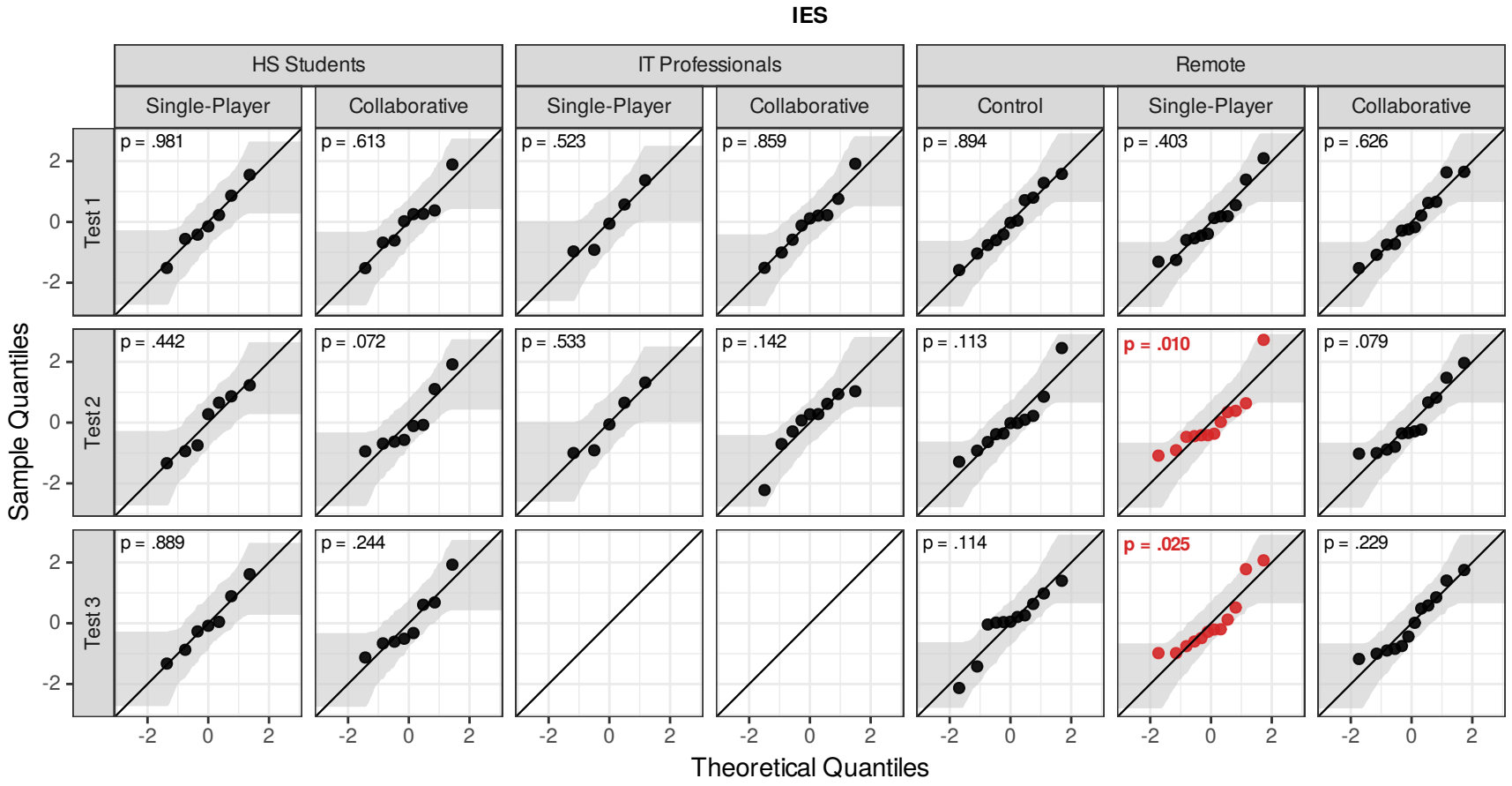


Figure C.15: Q–Q plots for IES by cohort and treatment condition

Appendix D

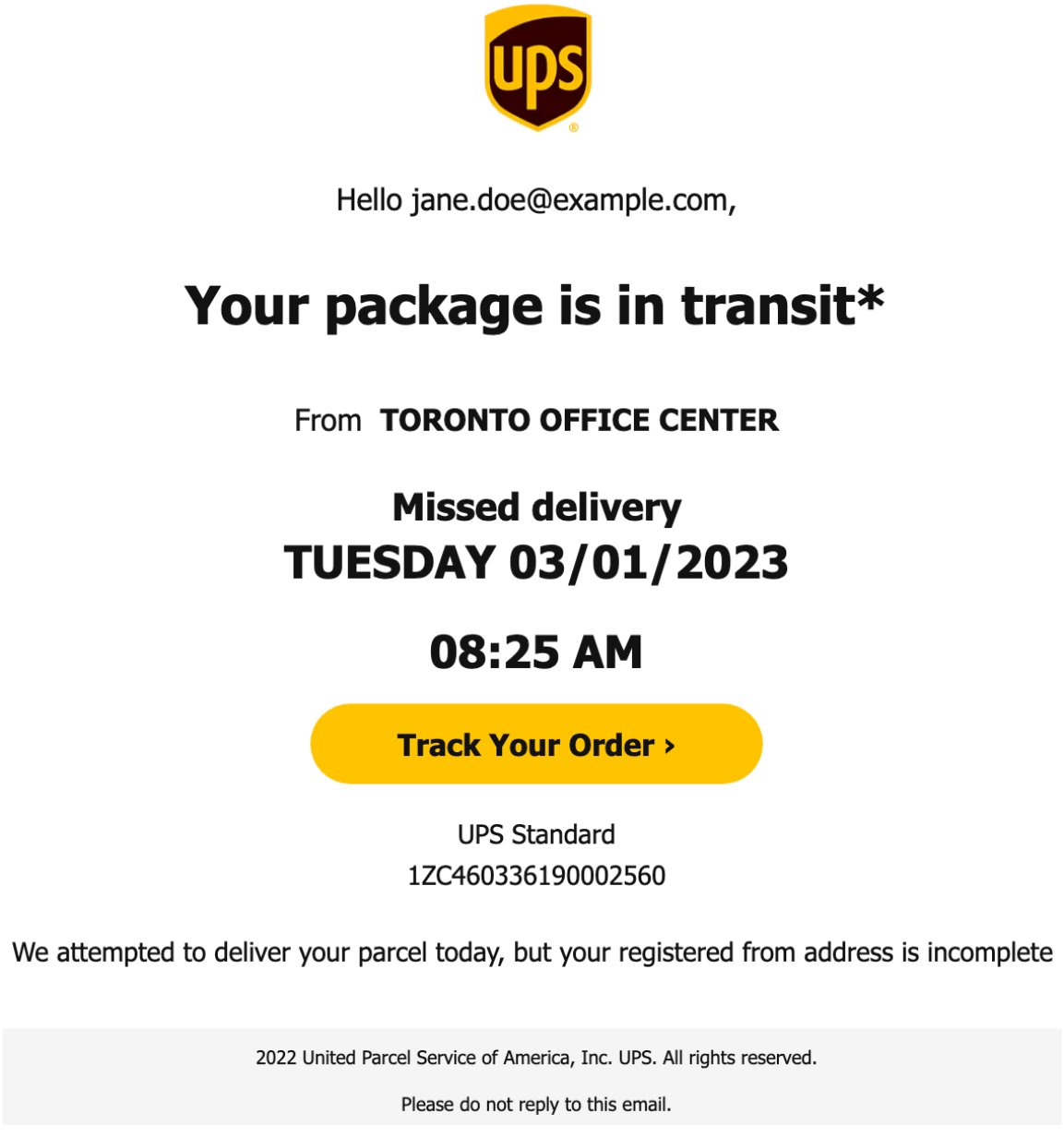
Sample Emails

This appendix displays twenty selected samples from the *Phooled!* database. Section D.1 contains Phishing samples, while Section D.2 contains Safe samples. For each sample, the metadata shown to the player is displayed at the top, followed by a screenshot of the rendered email. Samples that contain multiple HTML elements have their phishing cues highlighted inline where applicable, while those that contain only an image are displayed as-is. Any PII present in the email is replaced with the details of a generic user, namely Jane Doe `<jane.doe@example.com>`.

D.1 Phishing Samples

From: U.P.S <nakagawa@nomad.or.jp>

Subject: UPS Shipment Notification / Notification dexpedition UPS (430336190002530)



From: AppIe Customer Service <support@junsoft.com.br>

Subject: Your iCloud storage is full - Get more storage for free

---

Your iCloud memory is full

FULL

Hello Jane,

Your iCloud storage is full. Because you've exceeded your storage plan, your documents, contacts, and device data are no longer backing up to iCloud and your photos and videos are not uploading to iCloud Photos. iCloud Drive and iCloud-enabled apps are not updating across your devices.

To continue using these iCloud services, you need to upgrade to iCloud+ or reduce the amount of storage you are using.

Just for limited time (Giveaway ends tomorrow at 11:59PM)
Get NOW additional 50GB.

The iCloud Team

iCloud

iCloud is a service provided by Apple.
Apple ID | Support

From: RBC Royal Bank <f.morgan@gki.com>

Subject: Security Alert: Your Account Needs Attention

Your password has been disabled!

Your password has been entered incorrectly three times and for your security, we have disabled your password.

To enable your password, please sign in and follow instructions to review your recent activities.

Sign in to my account

We take your security seriously want to keep you in Me loop with important activities in your account. [Spelling/Grammar] So, if something does. look right or you weren't aware of Mese Manges, we are here to help.

Thank you for being a valued customer.

Enjoy value, flexibility, and choice

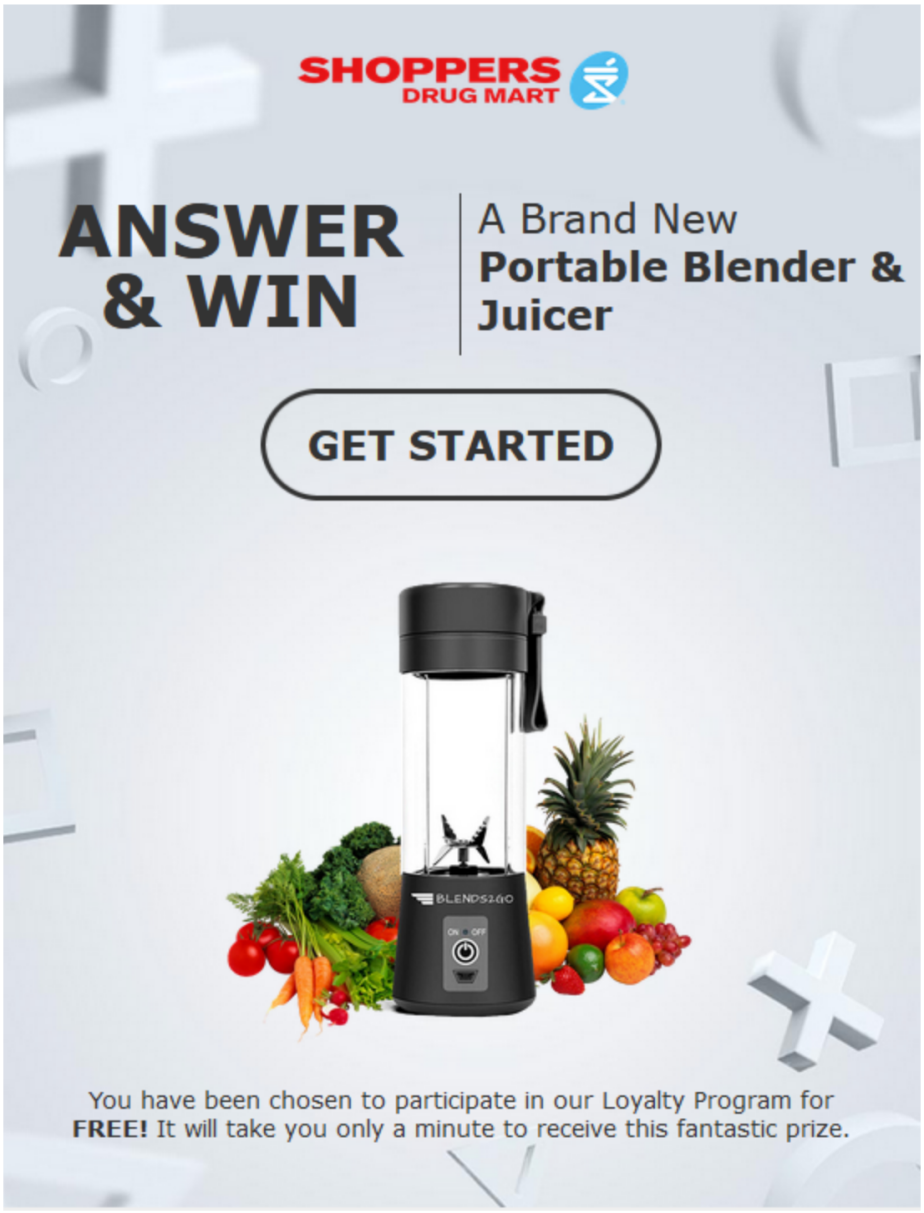
Are you signed up for RBC ULTIMATE REWARDS?

Here are just some of the perks you could get with RBC Ultimate Rewards.

From: Drug_Mart_support_17 <notification@walletib.app.br>

Subject: Re:Confirmation_Portable_Blender_70740

Welcome to Our Team Final notice for Portable Blender & Juicer



From: Delilah Brian <delilahbrian12@gmail.com>

Subject: RESEARCH ANNOUNCEMENT FOR STUDENTS

INTERNSHIP AT (UCLA)!!!

The Department of Information Studies at UCLA urgently requires the services of graduate/undergraduate students [Urgent Language] to work remotely as research assistants who will be paid $350 weekly. This position is open to both UCLA students and students from outside the university, regardless of their academic department.

The job is designed flexible and does not require a lot of prior experience. It is possible to accomplish tasks from the comfort of one's current location without interfering with their academic performance or daily activities.

To proceed with the application process, feel free to contact Christine L. Borgman at +1 (703) 769-7674 via text message stating your full name, email address, year of study, and department to receive the job description and further application requirements.

Best regards,

Christine L. Borgman
Title: Professor
Department of Information Studies
UCLA
Los Angeles,CA 90095,
United States
Phone:+1 (703) 769-7674

From: Qbooks_Update <jeangrey8823@icloud.com>

Subject: You're Officially Subscribed! Payment Received.

QuickBooks

Dear Customer,

Your Business Essential Plan subscription has been renewed and updated successfully. The charged amount will appear on your account statement within 24 to 48 hours. [Urgent Language] If you did not authorize this transaction, **please call us immediately on our customer support team +1 (844) 897-8810.**

User Id:	jane.doe@example.com
Renewal Date:	December 06, 2023
Payment Method:	Auto Debited
Order Id:	QB3442FN162446

Subscription Details:

Description	Tenure	QTY		Amount
Business Essential Plan	1years	1		$680.98
			Tax:	$60.00
			Total:	$740.98
			Discount:	**30%**
			Subtotal:	**$518.68**

Note:

This email confirmed that you've renewed your 1year subscription to QuickBooks for $518.68 on December 06, 2023. This subscription will auto renew every 1 year unless you turn it off, no later than 24 hours before the end of subscription Period. To cancel the subscription, you can reach us on +1 (844) 897-8810.

Intuit. | simplify the business of line

From: memberassistance@oas.ca <wmpo_help@wemakeprice.com>

Subject: OAS Update: Your Statement is Ready

Your Updated old Age Security Statement

- We've revamped your OAS statement for clarity.
- Please review annually to:
    1. Ensure your earnings record is accurate.
    2. View your estimated retirement benefits between ages 62 and 87.
- Get extra details with our fact sheets tailored to your age and earnings.

View Statement

- Contact us
- Find an office
- About OAS
- ID: #SSAYNQ4820517396

From: Admin IT HelpDesk <info@ueaviation.sbs>

Subject: ncf.ca Validation Service Alert 6/25/2023

Unusual login attempts on your ncf.ca account

Hello [...],

We noticed some unusual login activity with your account.

Email	jane.doe@example.com
Time	6/22/2023 9:04:51 a.m. (UTC)
Location	Singapore, Singapore
Device	Windows 10
Browser	Chrome
IP address	61.13.16.238

To make sure your account is secure, let us know if this was you.

This wasn't me

This was me

Why am I getting this email?

This message was sent to you by cPanel Cloud
6/22/2023 9:04:51 a.m.

From: Free Prime Subscription `<Nagaoka@s-parkle.co.jp>`

Subject: Your Prime Membership Free Subscription needs attention.

---

amazon prime

Dear jane.doe@example.com,

Your Amazon Prime Subscription no° 75737179706 is about to expire*

Get Started™ ››

This email was sent from a notification-only address that can't accept incoming email. Please don't reply to this message.

Cher jane.doe@example.com,

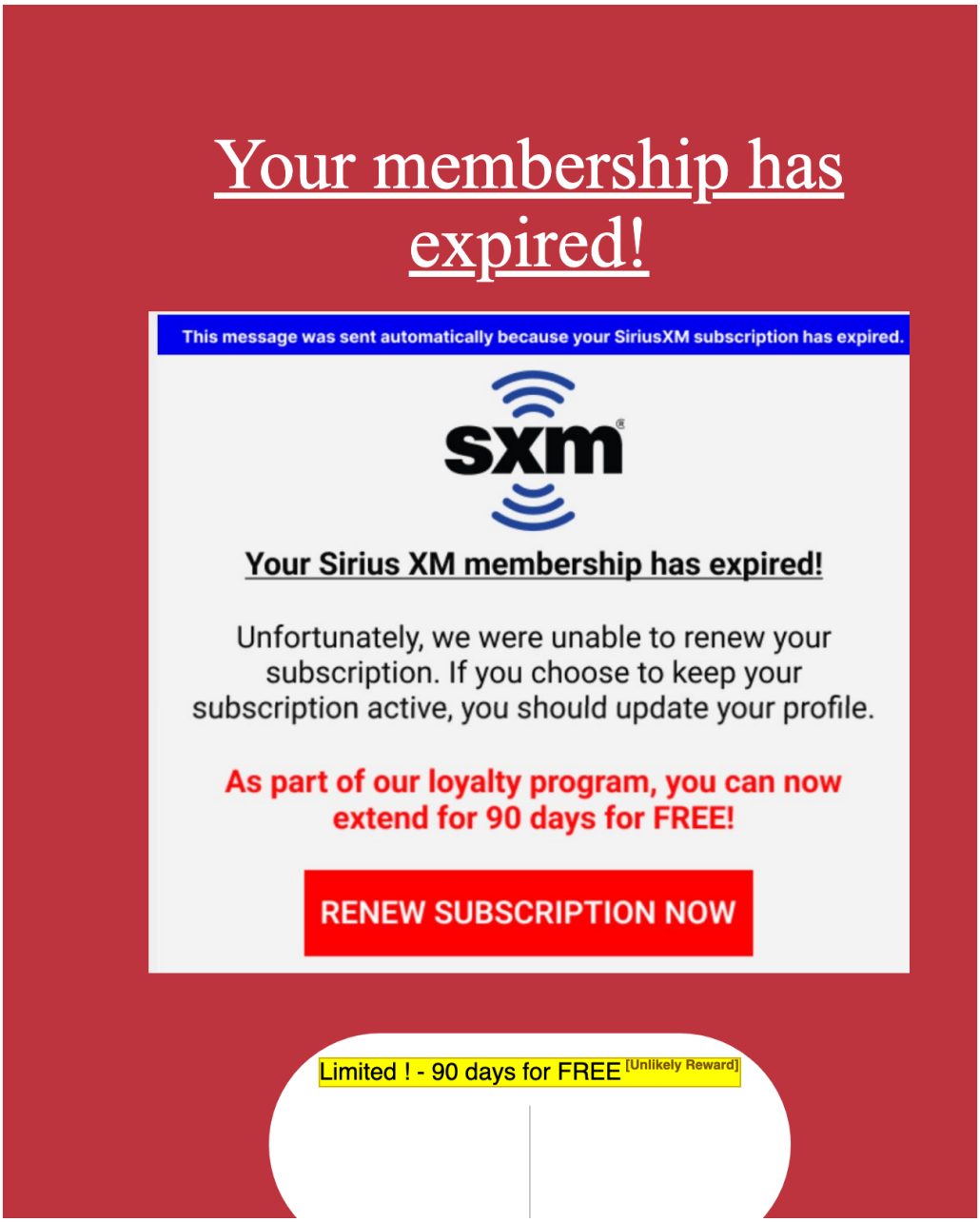
Votre abonnement Amazon Prime no°75737179706 est sur le point d'expirer*

Commencer™ ››

Cet e-mail a été envoyé à partir d'une adresse de notification uniquement qui ne peut pas accepter les e-mails entrants. Ne pas répondre à ce message.

From: Member Services <caekzn@mweb.co.za>

Subject: Your membership has expired!



Screenshot truncated below this point...

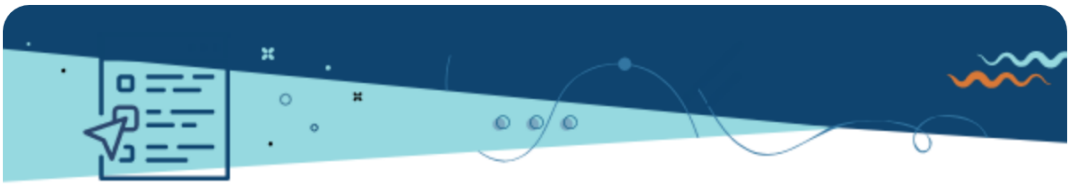
D.2 Safe Samples

From: Opinion Outpost <no-reply@opinionoutpost.ca>

Subject: Act fast! Your latest surveys are in

---

opinionoutpost



Time to start earning [Money]

You've got new surveys available, Jane.
Earn points for every survey you complete.

TAKE A SURVEY

This email was sent to: jane.doe@example.com.
To make sure you always receive our emails, add no-reply@opinionoutpost.ca to your safe senders list.
This is an automated email, please do not reply to this message.

From: Swagbucks <no-reply@swagbucks.com>

Subject: Swagbucks: Rated #1 Money Making site by MoneySavingExpert

---

swagbucks

Welcome & Congratulations Jane!

You just joined Swagbucks and qualify for a **juicy welcome bonus**!

Activate your account to get started or click the button below...

Activate My Account

(Activating your account also unlocks new earning opportunities. It's just one reason we're rated #1 Money making site by MoneySavingExpert.com!) [Money]

Sneak Peak: Members love rewards like these:

Have questions? Contact the Swagbucks Help Centre or see our Terms & Conditions.

From: Jennifer at Childhood Cancer Canada <info@childhoodcancer.ca>

Subject: I am a survivor who is pursing my dreams

Donate

Dear Jane,

I was four when I was diagnosed with Ganglioneuroblastoma (an intermediate tumor that arises from nerve tissues). The cancer was treatable after surgery, but it changed my life in so many ways. I've since been treating NF1 (neurofibromatosis type one), a genetic condition that affects my skin, eyes, and nerves, causing spots, tumours, and freckles. I also had an optic glioma (a slow-growing tumour) behind my eye.

Frequent hospital visits and out-of-pocket medications are expensive each month, and I often wondered if I'd be able to pursue my dream of attending university due to the cost of tuition. But Childhood Cancer Canada showed me that there is life beyond cancer, and I can dream big.

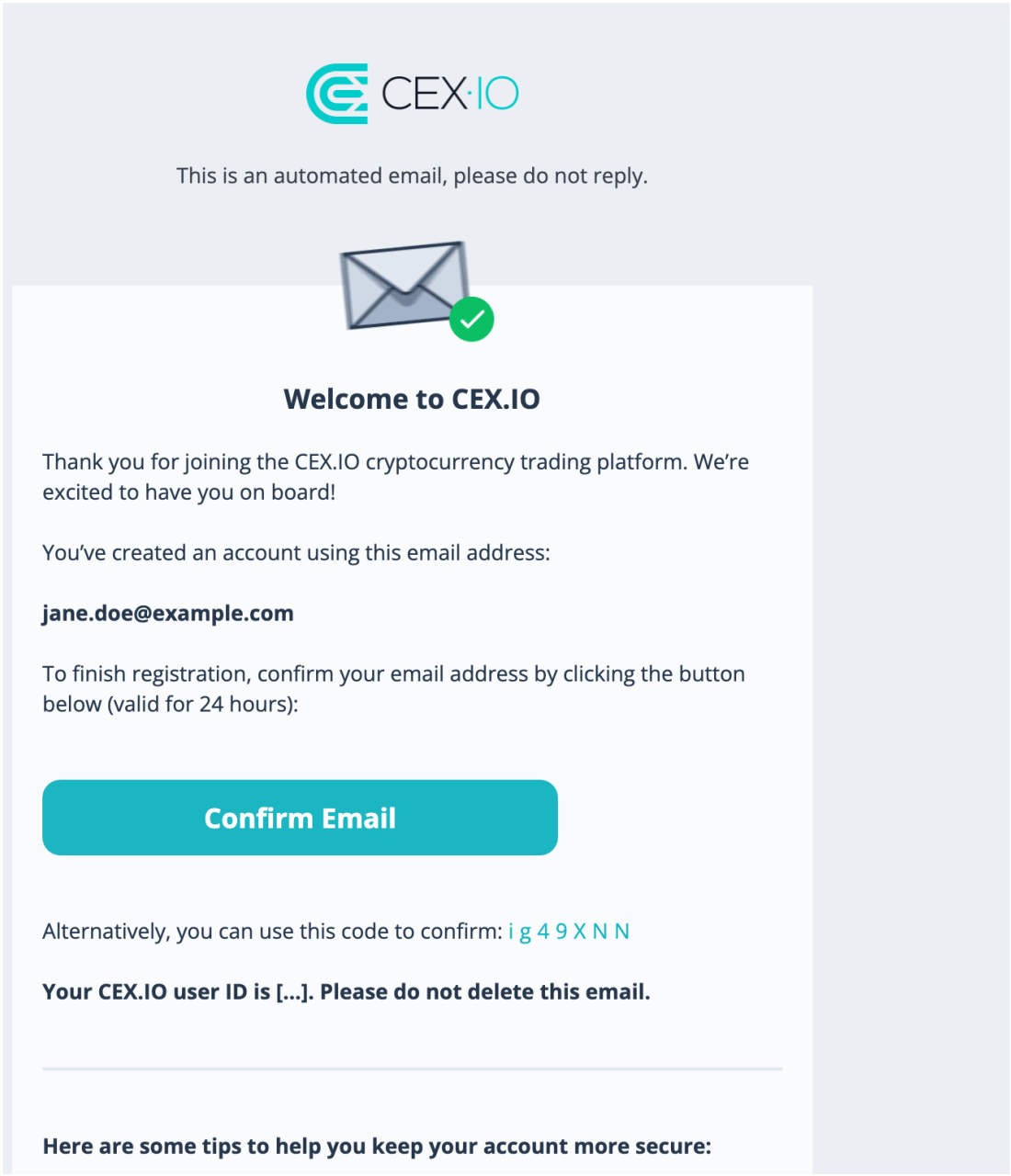
I received a Survivor Scholarship thanks to individuals like you who support Childhood Cancer Canada. Because of passionate people like you, I'm attending University and working towards my dream of pursuing a career in fundraising and marketing to one day help advance cancer research—it's my way of giving back.

Childhood Cancer Canada awards over 130 students each year with a scholarship but has nearly 170 applicants annually. The funds right now can only go so far. By making a donation today, you can help other young people pursue their dreams despite a cancer diagnosis. Every $1,500 raised means another cancer survivor can pursue higher education. [Money]

Screenshot truncated below this point. . .

From: CEX.IO `<noreply@cex.io>`

Subject: Confirm your email address



Screenshot truncated below this point...

From: BoardGameGeek (NO REPLY) <no-reply@mail.boardgamegeek.com>

Subject: BoardGameGeek Email Verification

TO VERIFY 'jane.doe@example.com' as a valid email address for the BoardGameGeek userid "Jane", click on the following link or enter it in your browser.

https://boardgamegeek.com/geekaccount/verifyemail?code=1f40f8c6&username=Jane

You may be asked to sign in first.

CODE: 1f40f8c6

Regards,
BoardGameGeek Admins

From: Rachel <natale@clarixevents.com>
Subject: Adobe Connect July Release

NEW | July Release of Adobe Connect

Discover unique features only available in Adobe Connect

Join the Adobe Connect experts at Clarix Technologies, inc. as we cover the **Adobe Connect June release feature set**. See all the BIG changes live along with the many added enhancements that allows Adobe Connect to stand out of the crowd!

Highlighted Features

Full Browser Experience: Join with just your browser, no plugins, no downloads
Persistent Rooms: Your room exactly how you left it, create and set your layout
Native Quizzing: Build and deliver assessments right in your virtual room
Customizable Events: Build and customize your designs for mobile and desktop

Additional enhancements:

Screenshot truncated below this point. . .

From: joan@mrjournal.cn

Subject: International reputation survey of Chinas scientific technological and medical journals

---

Dear Professor,

We hope this email finds you well. We are happy to intive [Spelling/Grammar] you to participate in the "International Reputation Survey of China's Scientific Technological and Medical Journals".

The purpose of this survey is to evaluate the international reputation of China's scientific technological and medical (STM) journals, investigate the awareness of international scientific researchers on China's STM journals through data analytics, and compare them with influential international academic journals in terms of article type, author text structure, and citation distribution. It should take around 10 minutes to complete and will be open until April 15th.

Survey participants also have the added benefit of receiving a report on the results after its completion.

To participate in the survey, please click here

Thank you for your time and participation!

Best regards,
"International reputation survey of China's scientific technological and medical journals" Peking University Project Group

---

If the message content does not display properly, click here unsubscribe complaint

From: noreply@td.com

Subject: You have a new secure message

You have a new secure message.

Please log in to EasyWeb® Online banking, or use the TD app to view your new secure message. [Money]

To view the secure message in EasyWeb:

1. Log in to EasyWeb at td.com
2. Select “Unread messages” on the Accounts page

To view the secure message in the TD app:

1. Log in to the TD app at td.com
2. Select your account
3. Click on the Statements tab
4. Click on View Other Documents

This mailbox is not monitored, please do not reply.

From: Mitacs Training <info@email.mitacs.ca>

Subject: Unlock Your FREE Courses with Mitacs Training

View in browser

mitacs

Hi Jane

Welcome to the Mitacs Globalink Internship program!

As a Globalink participant, we are pleased to offer **you exclusive access to professional development resources through** Mitacs Training.

Master 5 Skills Valued By Top Employers 💡

Discover FREE instructor-led courses—bilingual and fully accessible—that will help you develop five core competencies in the workplace:

√ **Leadership and project management**
√ **Interpersonal skills**
√ **Communication**
√ **Professional and career fundamentals**
√ **Intrapreneurialism**

Enroll in one or more courses, **enhance your resume** with the help of **Canada's top-ranking professional experts**—and **share your certificate** with your network!

Screenshot truncated below this point...

From: Aakash from Colab <support@letscolab.online>

Subject: Announcing 2024 Computer Science Internships Opportunities

---

Hey Jane,

Its Aakash from Colab, hope you have been well since we last wrote to you!

The last time we mailed you about the internship opportunities, we came across a massive demand but only had a few slots available for the interviews. Given the demand, we are opening up a few more internship cohort slots for the next week.

At Colab, we're gearing up for our 2024 Technical and Business Internship cohort, and it's not your typical gig. We're gonna get you ready for Spring'24 internships and here's what you can expect
:🚀 8-week live training in high-demand skills.
🌟 Build valuable network with industry legends
👥 Join a tight-knit community of interns + professionals across the US.
💼 Score guaranteed paid internships in your field.
🤝 Collaborate on result-oriented projects to demonstrate your readiness

We are not about waiting on applications [Spelling/Grammar] – directly schedule an interview, and we'll take it from there. Sound good?

Schedule Interview One

Glad help out anywhere you need, just hit reply. Catch you on the flip side Syrine, Cheers!

Aakash Dave
Founder and CEO, Colab
www.letscolab.online | Connect on Linkedin

COLAB

Sent with 💜 from Colab HQ

P.S: This email might be automated, it'll still reach my inbox.

miss out on all info

SENT BY mailer lite