What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game

Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, Erik Andersen
Cornell University
Ithaca, New York
{zw385,zl279,rc668,ela63}@cornell.edu

ABSTRACT

Phishing attacks are a major problem, as evidenced by the DNC hackings during the 2016 US presidential election, in which staff were tricked into sharing passwords by fake Google security emails, granting access to confidential information. Vulnerabilities such as these are due in part to insufficient and tiresome user training in cybersecurity. Ideally, we would have more engaging training methods that teach cybersecurity in an active and entertaining way. To address this need, we introduce the game *What.Hack*, which not only teaches phishing concepts but also simulates actual phishing attacks in a role-playing game to encourage the player to practice defending themselves. Our user study shows that our game design is more engaging and effective in improving performance than a standard form of training and a competing training game design (which does not simulate phishing attempts through role-playing).

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**; • **Social and professional topics → Adult education**; • **Applied computing → Computer games**.

KEYWORDS

Anti-Phishing; Applied Game; Situated Learning

ACM Reference Format:
Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, Erik Andersen. 2019. What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300338

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300338

1 INTRODUCTION

Phishing is the act of deceiving people into divulging information or unintentionally installing malware on their computers by sending the victim(s) counterfeit emails [33]. These counterfeit emails work by misleading the victim into thinking they come from a legitimate source. For example, a phishing email can link to an imitation of the PayPal login screen. Victims who believe the link is legitimate will enter their login credentials to the fake site, unwittingly giving the hackers access to their PayPal account. In addition to financial gain, government-backed hackers may disrupt elections by phishing specific persons who are affiliated with powerful institutions. In the 2016 US election, John Podesta, the chairman of Hillary Clinton's campaign, clicked on the change password link in a phishing email intended to look like a Google warning [65]. His action immediately unlocked some or all of his emails to the hacker.

To repel phishing attacks, phishing defense technology has evolved rapidly. Recent automatic systems apply machine learning to classify phishing emails, but these automated approaches are not foolproof [10]. There remains a non-negligible probability of users receiving phishing emails and these users must decide whether an email in their inbox is phishing or legitimate.

Research [29] shows that hackers can effectively and efficiently target end users due to the public's general lack of awareness regarding information security. People who are prone to taking risks are more likely to be phished [55]. Even if people are aware of phishing, simply knowing does not provide useful strategies for identifying phishing attacks [30]. On top of this, phishing attacks often convey a sense of urgency or utilize threats to pressure the recipient into responding [40]. Due to these factors, people judge the legitimacy of incoming message by visual cues, which can be easily copied by phishers [37]. Therefore, user education is a vital approach to protect users against phishing. While many organizations provide materials on how to defend against phishing attacks, such as email bulletins or information security websites [21], studies [39] found that these kinds of materials only work if people keep paying attention to them.

To better engage learners and change user behaviour, several anti-phishing games have been proposed. Although evaluations of these games have demonstrated an improvement in their players' ability to identify phishing websites, existing games leave out email context that hackers often leverage to demand immediate attention and encourage rash decision making. A person who can quickly parse a URL might hurriedly click on the hyperlink syntax without hovering on it because it seems to be an urgent request from the boss. Moreover, these game designs are not particularly effective at teaching players how to detect combined phishing techniques. For example, some who have played those games might not fall for malicious URLs in a phishing email, but they might click on the malware attachment enclosed in the same email. Incorporating these combined phishing techniques into our game's design can lead to more engaging challenges and more practical knowledge.

To develop a comprehensive anti-phishing game, we designed *What.Hack* (pronounced what dot hack), an online simulation game that features an engaging sequence of puzzles. Each puzzle requires players to study the anti-phishing rules in a rulebook that help evaluate whether an email is legitimate or phishing. Phishing emails in the game are generated by templates collected from real phishing emails. Players need to carefully identify phishing emails or they will encounter a bad ending (e.g. loss of integrity in the game).

To analyze the effectiveness of *What.Hack*, we tested the impact of the game on players' ability to recognize the real incoming phishing emails that came from a database of emails collected by the authors' university. We compare it against the current non-gamified training that the university uses and a good competing anti-phishing game, *Anti-Phishing Phil* [56]. We found that *What.Hack* achieved a 36.7% improvement in players' correctness in identifying incoming phishing emails from pretest to posttest, but did not find a statistically significant improvement for the training materials or *Anti-Phishing Phil*, which indicates that our context-based approach is effective for training people to recognize and defend against phishing emails. Moreover, examination of the feedback from players and logs of their play activity shows that players found the game engaging.

2 RELATED WORK

Games for Learning

Educational games can improve learners' performance, especially knowledge that is best learned actively through experience rather than passively. For example, *Crystallize* [25], a 3D video game for learning the Japanese language, showed that situated learning can be effective and engaging. *Reduct* [15] demonstrated that novices can learn programming concepts by playing a game. This suggests that gamifying education is potentially beneficial.

Anti-phishing education often struggles to capture the interest of end users. Materials commonly used for cybersecurity training include notes, videos, and email bulletins. However, these materials are often not very engaging and separate the learning material from the context in which employees routinely apply this information (e.g. email clients). Staff interviews by Conway et al. [21] revealed a continued desire for engaging cybersecurity materials that tie into daily experiences and practices. An investigation into the current state of cybersecurity education in industry produced similar conclusions [53]. According to both reports, anti-phishing education is working and more people are aware of the concept of phishing, but more work needs to be done. Our goal is to replace training programs that typically emphasize readings on cybersecurity with a role-playing game that mimics the actual situation of being phished.

Simulating Situational Context through Role-playing

Using the role-playing approach to engage students and improve their learning transfer performance has been a well-known design strategy in gamified education. For example, *Quest Atlantis* [17] is a 3D virtual learning environment that allows students to work together to perform educational activities that are known as Quests. There are other successful role-playing task-solving simulation games designed for the purpose of science and ethics education [35, 54], engineering internships training [19], etc. These games indicate the potential for supporting increased levels of engagement and learning across different domains.

Research in learning theory also supports the idea of presenting information in context. The theory of situated learning [41] stipulates that "the potentialities for action cannot be fully described independently of the specific situation " [12]:6. Shaffer [54] drew upon situated learning literature on communities of practice [41] when introducing epistemic frames for supporting learning transfer. According to encoding specificity theory [60], recall is highest when the context in which something is learned is perceptually similar to the context in which it is used. As Gee suggested, games that engage players in authentic situated problem solving facilitates learning can be transferred out of the game [31]. Experiments on tutorials in games [11] also showed evidence for the efficacy of presenting information in context in the case of the protein-folding game *Foldit* [22].

Game-based Cybersecurity Training Designs Review

Recent surveys [48, 57, 59] have summarized the current state of game-based cybersecurity training designs. We categorized the list of games mentioned in these surveys and compared their game type, target audience and design objectives with *What.Hack* in Table 1. We excluded cybersecurity games for children [1, 5, 8, 34] because they have a slightly different

Game Type & Examples	Description	System Attack	URL Phishing	Spear Phishing
Board Games [28, 32, 49, 57, 62]	Teach high-level security concepts.	✓	✓	✓
Capture-The-Flag [3, 16, 18, 47, 52, 64]	Let coders compete for scores by defending their systems and hacking others'.	✓		
Sys-Attack Sim RPG [2, 6, 20, 58]	Teach players to defend against computer system attacks in a realistic system attacks simulation game.	✓		
Non-Phishing RPG [14, 44, 56, 67]	Teach players to identify phishing URLs in a cartoon-like game without phishing attempts.		✓	
Phishing Sim RPG *What.Hack*	Teach players to defend against URL and spear phishing attempts in a realistic phishing simulation game.		✓	✓

Table 1: Game-based Cybersecurity Training Designs Comparison

game design goal. In general, these games focus on introducing basic tips of staying safe online. Tips for kids are easier to practice in general and some might even be considered inapplicable under most real-world situations such as a corporate workplace. For instance, Carnegie Cadets [1] suggests that one should look at the sender's username and the subject title to identify spam emails, which is a habit that is actually exploited by some phishing attacks. In the following subsections, we discuss each of the game types listed in Table 1.

Board Games. Control-Alt-Hack [28] is a board game that teaches players high-level security concepts such as phishing, social engineering, etc. While this does help to increase awareness and understanding of cybersecurity topics as a whole, it is not sufficiently specific enough to simulate the low-level decisions required for anti-phishing strategies in practical contexts. In general, board games in information security [32, 49, 57, 62] are not meant to teach hands-on security skills, such as how to identify phishing attacks, which are the main focus in our game.

Capture-The-Flag. Capture-The-Flag (CTF) is a game-based computer security competition for students to practice skills of defending against hackers. There are two types of cybersecurity CTF: attack-defend and Jeopardy-style [27]. In attack-defend CTF, each team attacks other teams' servers and protects their own server. The "flags" are files in the defending computer that the attack team attempts to retrieve as they compromise the computer. In Jeopardy-style CTF, teams solve puzzles by using knowledge like cryptography, coding, etc. Solving a puzzle means they capture a flag. The team with highest number of flags wins. Researchers [27, 43, 58] have shown that CTF is an engaging and effective way to motivate coders to learn how hackers think and how to defend against them. However, these competitions usually require basic coding background to learn about hacking techniques. Therefore, they are not directly applicable for teaching the general public about anti-phishing techniques.

System-Attack Simulation Role-playing Game. To engage the general public to learn how to defend a computer system or network settings from being compromised, CyberCIEGE [58] and similar approaches [2, 6, 20] adopted an attack simulation role-playing design that is similar to our game but focuses instead on system attacks. A user study found CyberCIEGE to be engaging even when students knowingly fail, but did not assess learning effectiveness [58]. In this work, we further examine whether the potential for increased effectiveness and self-efficacy in handling real-world cyber threats justifies the effort of designing a role-playing game to simulate situated phishing attempts.

Non-Phishing Role-playing Games. Recently, new designs for anti-phishing games [14, 44] have drawn inspiration from the popular game design framework for anti-phishing training [13] and *Anti-Phishing Phil* [56], which do not simulate situated phishing attempts. *Anti-Phishing Phil* is a representative Non-Phishing RPG that teaches players how to identify phishing URLs. In each round, players act as a fish to "eat" the worm that shows safe URLs and "reject" the bait that shows phishing URLs. Visualizing URLs as worms makes the game fun, but it also takes the URLs out of context and does not simulate the real experience of detecting phishing emails.

The authors of *Anti-Phishing Phil* mention two limitations. One is their focus on URLs and domain names, which are only two of the phishing attack templates, leaving the player vulnerable to content-based attacks like spear phishing attacks. The other limitation is their exclusive focus on URL syntax without addressing URL semantics, meaning their players are still vulnerable to URL semantics attacks (safe URL syntax redirects to forged URL address). These limitations also apply to recent games [14, 44] that are extended from it. One of

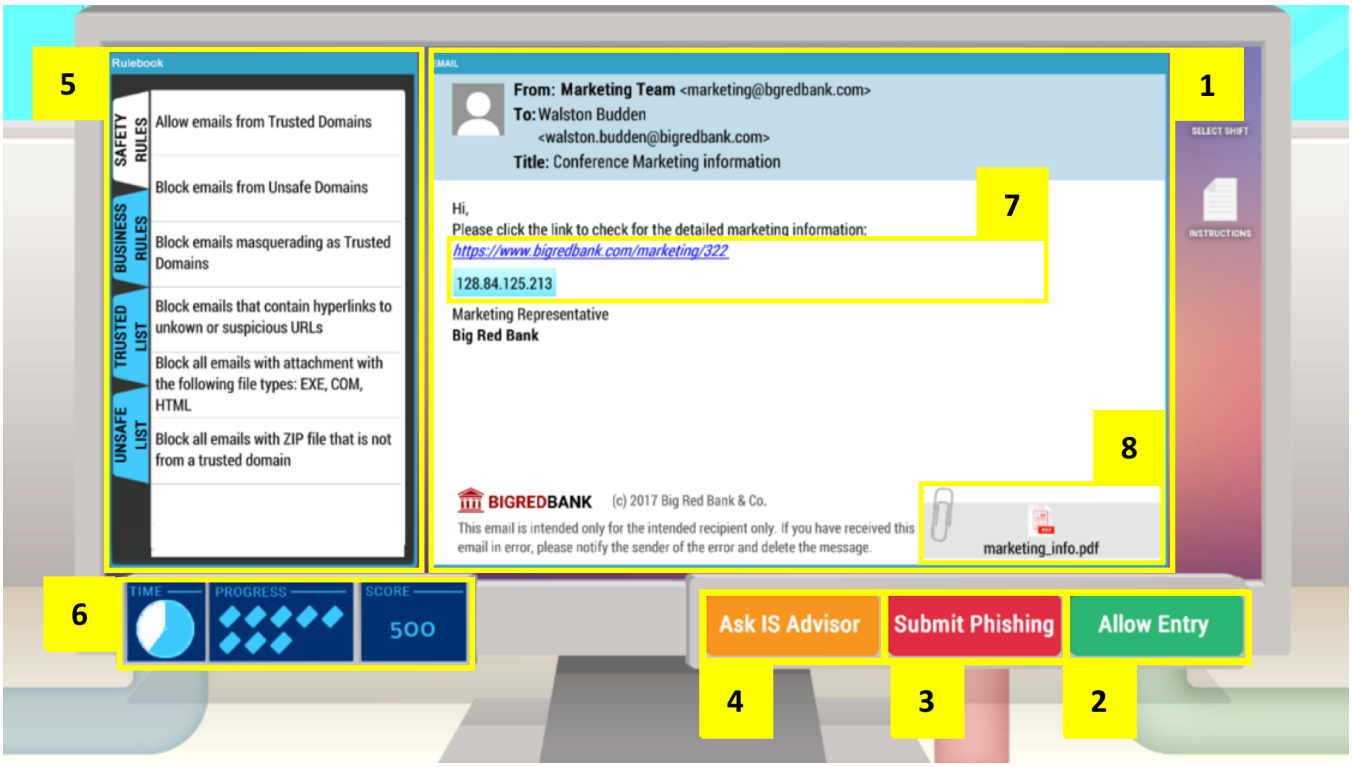

Figure 1: In *What.Hack* players process emails to acquire contracts and to protect their network from cyber criminals (1). Player either select *Allow Entry* button to let the email reach the recipient (2) or select *Submit Phishing* button to block it from doing potential harm (3). If players cannot determine whether the email is dangerous, they can click *Ask IS Advisor* to take some time to analyze the email (4). Players can refer to the rulebook to decide whether this email is legitimate or phishing (5). Players need to process a number of emails within a time limit (6).

our main goals with *What.Hack* is to achieve more holistic anti-phishing education.

An anti-phishing game design framework [13] includes design elements such as safeguard effectiveness and perceived susceptibility. Multiple Non-Phishing RPGs [14, 44, 67] as well as *What.Hack* address all of these elements. That said, we argue that this framework is sound but does not sufficiently address the identification of real-world phishing threats if design elements are not implemented in a context that resembles the one used by hackers. Evidence from a psychological study [50] showed that certain types of phishing emails, such emails that sound like they come from a friend or warn of some kind of failure, are more effective than other types, such as emails that offer a deal, regardless of the specific phishing templates the hackers use. Our goal is to implement phishing attempts in a way that closely matches how they occur in the real world.

3 GAMEPLAY DESIGN

We have developed a prototype game, *What.Hack*, which facilitates role-playing for learning email phishing defense. A screenshot of *What.Hack* is shown in Figure 1. *What.Hack* was designed with three primary learning goals: 1) teach email phishing defense in context by replicating as many real-life conditions as possible, 2) engage the player by setting clear goals and tasks that become more difficult over time, 3) provide immediate feedback about the consequences of decisions the player makes.

What.Hack attempts to teach phishing emails defense through the game mechanics themselves. The core game mechanics encourage players to screen incoming emails to determine whether they might be malicious, and provide a set of constraints ("rules") that players can use to evaluate an email. Over time, the rules become more specific and combine to create a complex rule set, simulating the large amount of constraints that must be applied in real life.

The key challenge in constructing situated learning approaches for anti-phishing centers on how to replicate the perceptual characteristics of the situations in which real people defend against phishing attacks in a way that is fun, educational, and not tedious. To do this, we designed a training module that provides rich visual and situational context. In

What.Hack, the player assumes the role of an employee working for a bank. To win, players need to help their bank acquire contracts by processing business-related emails and avoiding being phished at the same time.

Porting a Subset of *Papers, Please* Game Mechanics to a Phishing Simulation Game

Since processing emails is not generally considered compelling, we investigated existing games with game mechanics centered around document inspection. We observed that *Papers, Please* [42], a popular and highly engaging game released in 2013, focused heavily on document inspection. In this game, the player acts as a border patrol officer. The player must review the passports and other supporting documents of entrants seeking admission to a fictional communist country. To do this, the player must apply a set of rules specified by the border control office, which are expressed to the player through a rulebook. The player is directed to accept only passports that come with valid paperwork and reject or detain those with improper forms.

While designing *What.Hack*, we adapted some of these document inspection mechanics to the anti-phishing context of processing business emails to foster motivation for learning email safety rules. In our game, players can allow their colleagues' business emails to go through or to block them, which will affect their company's success. Learning email safety rules becomes necessary for players to make progress in the game. With the shift of purpose to anti-phishing training, players must now prevent their company from being hacked and help it to prosper.

Papers, Please includes a range of game mechanics that we did not adopt in *What.Hack*. *Papers, Please* makes a critique of Orwellian communist bureaucracy and dehumanizing processes [45]. In addition to the narrative, the game introduces several ethical dilemmas, forcing the player to choose between obeying oppressive laws and taking risks to advance moral causes. Since we did not incorporate critiques or ethical dilemmas into *What.Hack*, a significant portion of what makes *Papers, Please* compelling and unique was lost in this conversion. However, we consider this acceptable since our primary goal was to adapt the document inspection mechanics for anti-phishing, not to raise questions about the nature of the organization in which the player operates.

Simulating Email Processing Context

The following sections highlight four different ways in which *What.Hack* simulates the real-world context of email processing: 1) workflows, 2) time pressure, 3) interactions with IT support and 4) harmful effects of phishing.

Simulating email processing workflows. A major goal in our game design is to simulate the real-life trade-offs in processing incoming emails. Although deleting every incoming email is a solid defense against phishing attacks, this is implausible in many situations because legitimate important emails will be missed. Therefore, in *What.Hack*, the player must consider each email and decide whether to accept or reject it. The player must make this decision without performing a dangerous action (such as opening an email attachment that could potentially contain a virus).

The player interacts with a simulated operating system, as shown in Figure 1. From this screen, the player can launch other application windows. The application that plays the biggest role in the game is the email client, in which the player can view a stream of emails that have been sent to the player. Some of these emails are related to legitimate business interests of the bank, and others are not. Some emails are malicious phishing attempts.

The player indicates their intention to accept or reject an email by clicking on the buttons in the lower right of the screen. If the email is business related and safe then the player should select the "Allow Entry" button (highlight (2) in Figure 1). Otherwise, the player should select the "Submit Phishing" button (highlight (3) in Figure 1) to "shred" the email. If the player "shreds" too many legitimate emails, the player's manager will show up and inform the player that they underperformed because their customers complained that the player is unresponsive. This mechanism mimics the real-life decision process in which every employee has to decide whether to trust an incoming email and figure out the appropriate response.

Simulating time pressure. A qualitative investigation by Conway et al. [21] indicated that employees are more vulnerable to phishing links and attachments when they are swamped by a large amount of work. Due to time pressure, they need to scan incoming emails and react quickly if the email is relevant to their work so that they can finish tasks in time. Therefore, it is critical to simulate the time pressure that phishing victims often experience. This is implemented through a time limit, indicated by highlight (6) in Figure 1. The game is organized into five "shifts," or levels, which require the player to process an average of six emails in eighty seconds. Therefore, if the player takes too long to consider each email, the player will not pass the shift. This also incentivizes the player to work fast - the player will obtain a higher score if the player can accurately process more emails than required.

Simulating interactions with Information Technology support. Since phishing remains such a prevalent problem, many IT departments have developed mechanisms for staff to report phishing emails and they encourage people to do so. This can sometimes help prevent other people from falling victim to the same email. Since this is a key component of many people's email processing workflow, we simulated it in our game.

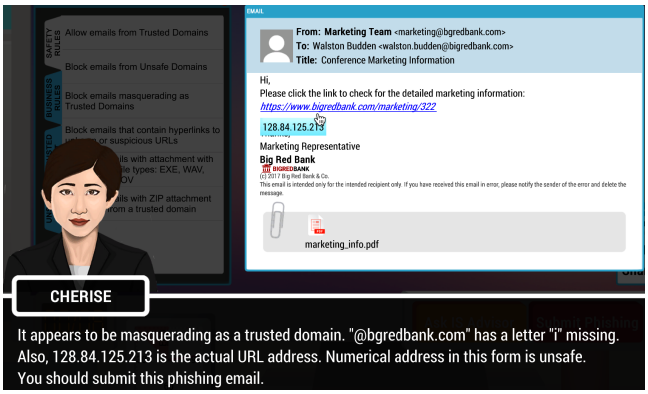

Figure 2: After clicking on the "Ask IS Advisor" button, Cherise, the information security advisor, appears and says: "It appears to be masquerading as a trusted domain. "@bgredbank.com" has a letter "i" missing. Also, 128.84.125.213 is the actual URL address. Numerical address in this form is unsafe. You should submit this phishing email.". The corresponding email is highlighted.

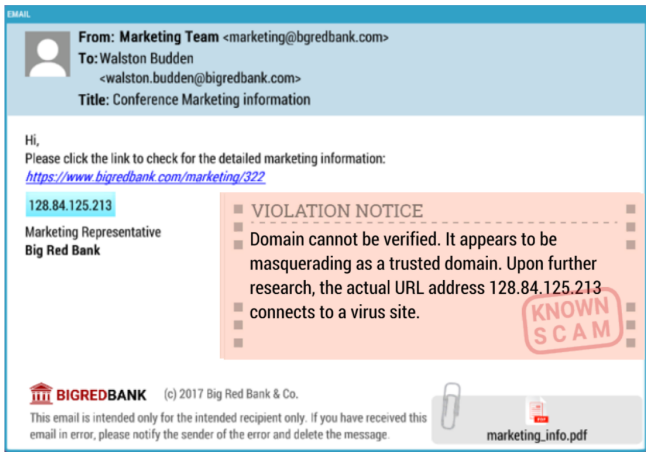

Figure 3: *What.Hack* provides immediate feedback when the player makes a decision. Here, the player clicked "Allow Entry" for an email that contained a malicious link. The player has now received a violation explaining the mistake. Violations affect the player's overall progress and score and acquiring too many violations in a shift will result in the player being required to repeat that shift.

The button for *rejecting* an email is called "Submit Phishing" (highlight (3) in Figure 1).

Furthermore, many IT departments will provide consultation on whether or not an email is safe. However, this takes time, creating a tradeoff for staff members who experience pressure to process many emails quickly. Therefore, we implemented this tradeoff through the addition of a button that lets the player refer the email to the Information Security (IS) advisor for help if they are uncertain whether email is safe. After players click on the "Ask IS Advisor" button (highlight (4) in Figure 1), the cybersecurity advisor will appear and tell the player whether this email is safe or unsafe (shown as in Figure 2). If the email is unsafe, she will also explain the reason. A small amount of limited game time (2-4 seconds) are deducted to simulate the cost of communication. This mechanism encourages player to think actively whether this email is a phishing email before asking for help.

Simulating Harmful Effects of Phishing. A big challenge in cybersecurity training is teaching learners to appreciate the risks of a bad decision. Existing training materials often state these risks at a high level but do not really *show* the learner what will happen if something goes wrong. Therefore, *What.Hack* simulates how the player's decisions lead to various outcomes, both positive and negative. From the standpoint of maximizing engagement, a key consideration is how to give the player sufficient *freedom* to make meaningful decisions that will have both short-term and long-term consequences on the game, while still ensuring that the players understand what they do is right or wrong.

If the player makes a wrong decision, a violation note will appear, as shown in Figure 3. If the player labels an unsafe phishing email as safe, the note will state the specific rules that the email violates. Similarly, if the player thinks a safe email is a phishing email then the note will remind the player that this email is safe. This approach helps players reflect on unfamiliar anti-phishing knowledge, which helps retain knowledge [23]. If the player fails too much, the player's bank loses trust and the player gets fired.

Structuring the learning content

Our goal for the level design of this first *What.Hack* prototype was to measurably improve the player's ability to recognize potentially malicious emails within a short amount of play time. Therefore, we focused on three popular phishing attack templates [53]: 1) similar domain attack, 2) URL manipulation, and 3) malicious attachment.

In order to maximize engagement, we constructed the shifts so that they combine concepts in order to form a progression that starts easy and grows more difficult. The progression design motivated by the theories of flow [24], elaboration [51], and the Zone of Proximal Development [61]. This contrasts with existing anti-phishing training games that typically focus on only one concept at a time, and only combine concepts on the last level [14, 56].

The progression generally increases one or two attack templates per shift (level), and continually combines these attack templates with other templates introduced in previous shifts.

The similar domain attack is first introduced at Shift 1. The URL manipulation attack happens at Shift 3; and the malicious attachment attack appears at Shift 4.

Balancing Freedom and Learning through the Rulebook. Traditional email embedded training often relies on absolute rules like "never click on links", which are impractical in real life. To improve, the rules taught by our game should be meaningful rather than purely prescriptive. Furthermore, we designed the game so that players can perform actions at will but also learn which actions lead to negative consequences. A key innovation of *Papers, Please* is that the target skills (related to border control document verification) are communicated to the player through a *rulebook*, to which rules are gradually added over the game. In *What.Hack*, we explored whether the same rulebook mechanism could be used to deliver learning content related to phishing defense. The rulebook is shown in highlight (5) of Figure 1. It can be viewed at will, thereby serving as reference material for the learning progression.

Figure 4 shows an example of how a player might use the rulebook to determine whether an incoming email is malicious. Table 2 shows how the rules are introduced in the learning progression.

Shift 1	Allow emails from Trusted domains Block emails from Unsafe domains Block emails masquerading as Trusted domains
Shift 2	Block any email with inappropriate content Block non-business related emails from unknown domains Allow business-related emails from unknown domains
Shift 3	Block emails that contain hyperlinks to unknown or suspicious URLs
Shift 4	Block all emails with attachment with the following file types: EXE, COM and HTML Block all emails with ZIP attachment that is not from a trusted domain
Shift 5	Block all emails with DOC attachment that is not from a trusted domain but DOCX attachment is acceptable Ask IS advisor about emails from unknown domains that has a business-related attachment

Table 2: The rules that are introduced in each shift. The ruleset for each shift gradually grows in complexity, requiring players to make more realistic decisions in later shifts.

4 PERFORMANCE EVALUATION

Existing anti-phishing games that do not simulate phishing attempts through role-playing have shown that they were more engaging and thus provided better learning outcomes than non-gamified training [14, 56]. However, as pointed out by the criticisms of gamified education [46], not all gamified approaches are equally effective. To the best of our knowledge, there is a real lack of user study comparing two different approaches to the same gamified problem. Therefore, we ran an in-lab study to compare our game with a representative non-phishing role-playing game in addition to a non-gamified training.

Study Design

In order to evaluate the effectiveness of *What.Hack*, we conducted user studies that measured the impact of the game on players' ability to recognize phishing attacks on a pretest and a posttest. In the process, we compared the game to two anti-phishing training approaches.

We compared our game with *Anti-Phishing Phil* [56], a competing non-phishing role-playing game that educates players to identify similar domain and URL manipulation attacks, which has been cited for over 300 times. Its latest version is also provided for cybersecurity training in Carnegie Mellon [7]. Its gameplay covers the key elements that are addressed in the most popular game design framework for anti-phishing training [13]. The recent new designs [14, 44] of anti-phishing games followed the fashion of *Anti-Phishing Phil* gameplay. These evidences show that it is a representative anti-phishing game with which we decided to compare.

In addition to *Anti-Phishing Phil*, we also compared with *PhishLine* training materials [9] that are currently used by Cornell. *PhishLine* training materials are the typical form of fact-and-advice training that have been widely used and studied [36, 38, 63]. *PhishLine* and *What.Hack* cover three popular phishing attack templates: 1) similar domain attack, 2) URL link manipulation and 3) malicious attachment. *Anti-Phishing Phil* only covers the first two attack templates.

What.Hack targets young professionals and adults. We recruited target players through fliers, face-to-face interactions and Cornell's experiment sign-up system. Participants who signed up using the university's system received 2 experiment credits upon completion. We required the participants to be at least 18 years old and that they had never taken a cybersecurity class or participated in anti-phishing training. We recruited 39 students at Cornell and randomly assigned 13 people to each group.

The user studies consisted of an effectiveness evaluation session and an engagement evaluation session.

In the first evaluation session, which targeted effectiveness, participants were given a pretest in which they identified whether an example emails were phishing or legitimate. After the participants completed the pretest, we randomly assigned them to play *What.Hack* or play *Anti-Phishing Phil* or study *PhishLine* training materials with equal probability. All participants were able to finish their assigned game within half an hour. Participants were then given a posttest using the same emails in the pretest. The posttest also asked players to describe the knowledge that the participants thought they learned from the game that they played.

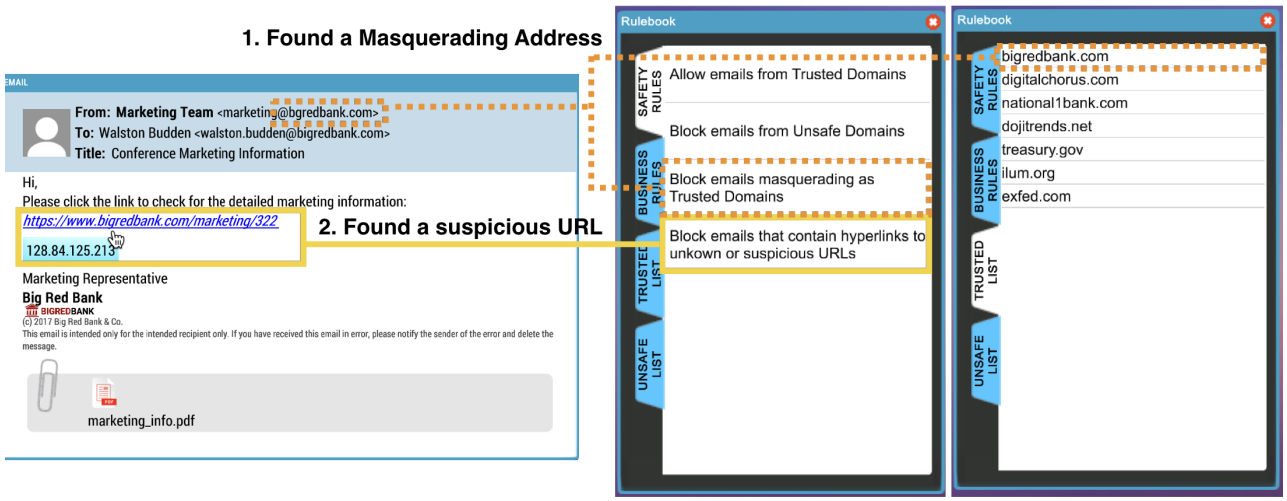

Figure 4: Players apply the rules from the rulebook to the email to determine if an email is a phishing attempt or not.

Attack Templates	URL Only	URL + Domain	Attachment Only	Attachment + Domain	Total
# of Phishing Emails	2	4	3	2	11

Table 3: Attack Templates in the Selected Phishing Emails for the User Study

In the second evaluation session, which targeted engagement, participants played the game that they did not play in the first session. After finishing the game, participants were asked to complete an exit survey about engagement.

Test Design

The goal was to confirm whether *What.Hack* can improve the correctness of identifying phishing emails with a statistically significant result. Therefore, we presented participants with emails in the pretest and the posttest. We asked them to decide whether an email is phishing or legitimate. We ask participants to choose whether it is phishing or legitimate. In addition, they need to rate how confident they are about the answer using the 5-point Likert scale from 1 (random guess) to 5 (very confident).

We selected and fine-tuned 11 real phishing emails from a database maintained by Cornell's IT security office, which is a similar approach that [63] took to select the first 12 phishing emails for their user study. The types of messages include suspicious account activity warning, financial document review, university president announcement, shipment notification, etc. Table 3 shows the number of phishing emails that use certain types of attack templates. These phishing emails can be seen in our online appendix: osf.io/pven9/.

In addition to the 11 phishing emails, we also chose 9 authentic emails from the communications database verified by Cornell's IT security office. They are not publicly available due to the privacy policy.

We examined the conceptual knowledge and procedural knowledge that participants retained. After they completed the quantitative test in the pretest and in the posttest, participants were asked to answer the question: "What is the strategy you used to process these emails? Please write in bullet points." This question allowed us to better understand how they make the decision before and after playing the game.

In addition to participants' strategy for identifying phishing emails, we asked them the following question at the end of the posttest to identify what else they had learned from the game:

"Did you learn any new concepts or skills from this training that will help you prevent yourself from being hacked by unsafe or phishing emails? Please write in bullet points."

To evaluate engagement, we asked participants the following 5-point Likert scale rating questions in the exit survey:

"On a scale from 1 (very boring) to 5 (very engaging), how would you rate the engagement of each training?"

"On a scale from 1 (strongly disagree) to 5 (strongly agree), how likely are you to recommend this training if your friends want to learn how to defend attacks from phishing emails? "

Results

We measured the effectiveness of our game by examining the correctness percentage, the false negative rate and false positive rate before and after the game. A false positive is

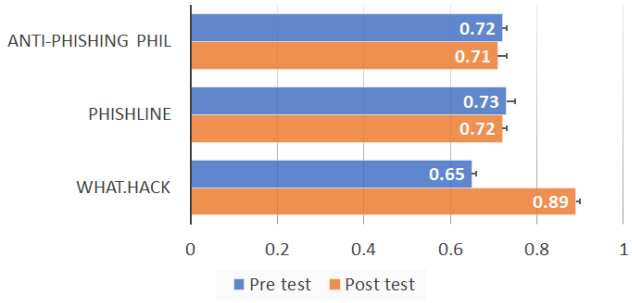

Figure 5: The correctness percentage. Our game improved players' correctness in identifying phishing emails by 36.7% **(statistically significant), whereas neither of the control groups achieved a statistically significant improvement.**

when a legitimate email is incorrectly regarded as a phishing email. A false negative is when a phishing email is incorrectly regarded as a legitimate email. In our case, the false negative rate is more important because it would expose people to the danger of phishing if they mistrusted the phishing email.

Evidence that the game improves correctness. We derive a measure for the change of correctness percentage between the pretest and the posttest using one-way ANOVA. There was a significant effect of training for the three conditions ($F(2, 36) = 18.53$, $p < .01$) with a large effect size ($\eta^2 = .507$).

Post hoc comparisons using the Tukey HSD test indicated that the score increase for *What.Hack* (M = .239, SD = .106) was significantly different than *Anti Phishing Phil* (M = -.004, SD = .107, $p < .01$) and different than *PhishLine* (M = -.007, SD = .138, $p < .01$). However, *Anti Phishing Phil* condition did not significantly differ from *PhishLine* condition ($p = .90$).

Evidence that the game enhances anti-phishing self-efficacy. The result shows that participants became more confident about their judgments after playing *What.Hack*. The average confidence rating increased from 3.33 (variance = .368) to 4.08 (variance = .329). The distributions between the two sets differed significantly (Mann-Whitney U = 32.5, Z = -2.64, $n_{pre} = n_{post} = 13$, p = .0083, two-tailed). While participants made a wrong decision after playing *What.Hack*, their confidence rates did not significantly change. This evidence meets our overall educational goal that players should be more confident of identifying phishing emails correctly while not making hasty decisions. We did not observe the control groups enhancing participants' confidence in a statistically significant way: the average confidence rating of *Anti-Phishing Phil* was 3.38 pretest (variance = .332) and 3.51 posttest (variance = .517); the average confidence rating of *PishLine* was 3.51 pretest (variance = .307) and 3.58 posttest (variance = .415).

Evidence that the phishing simulation game facilitates learning transfer. To compare each group of participants' ability to transfer the knowledge they have learned from the game or materials to identify whether an email contains malicious content, we removed 3 email examples that requires knowledge of attachment file types to make informative decisions because *Anti-Phishing Phil* does not teach how to identify malicious attachments. We run the analysis on the rest 17 emails that can be determined by validating the domain address and/or the URL hyperlink. The training games/materials still makes a statistically significant impact on the change of correctness percentage between the pretest and the posttest for the three conditions ($F(2, 36) = 13.34$, $p < .01$) with a large effect size ($\eta^2 = .426$) using one-way ANOVA.

Post hoc comparisons using the Tukey HSD test indicated that the score increase for *What.Hack* (M = .199, SD = .11) was still significantly different than *Anti-Phishing Phil* (M = -.013, SD = .12, $p < .01$) and different than *PhishLine* (M = -.02, SD = 0.14, $p < .01$). However, *Anti-Phishing Phil* did not significantly differ from *PhishLine* ($p = 0.90$).

Evidence that the game is engaging. There was a statistically significant difference between the engagement ratings of three training games/materials (Kruskal-Wallis H(2) = 44.121, $p < .01$), which a median rank of 4 for *What.Hack*, 3 for *Anti-Phishing Phil* and 3 for *PhishLine* (a score of 4 is positive, 3 is neutral). The median rank of rating of recommending three games/materials to friends who want to learn about defending phishing email were *What.Hack*: 4, *Anti-Phishing Phil*: 3, *PhishLine*: 3; the distributions in the three groups differed significantly (Kruskal-Wallis H(2) = 28.75, $p < .01$). Figure 6 and 7 are the bar charts of the engagement ratings and the recommendation ratings respectively.

Evidence that players learned new anti-phishing skills from What.Hack. We find evidence from the posttest questionnaire that some participants learned to defend the classic URL semantic phishing attacks after playing *What.Hack*: *'you need to hover your mouse to check the actual link'; 'the actual link will appear after hover my mouse on it'*. And some participants learned to not rely solely on the sender's identity or the content of the email to identify phishing emails: *'do not download exe even if it's forwarded by Cornell ppl'; 'the content is not most reliable way to identify phishing emails'*.

5 DISCUSSION

We believe that *What.Hack* is the first anti-phishing game that simulates phishing attempts through role-playing. Our evaluation indicated that *What.Hack* is effective and engaging, when compared to *Anti-Phishing Phil* and *PhishLine*. The results showed strong evidence that leveraging situated learning theory can also enhance anti-phishing training, in addition to domains such as science and ethics education [17, 35, 54].

Design implications. The user study participants thought that *Anti-Phishing Phil* was more engaging but less recommended than watching videos. The qualitative measures showed

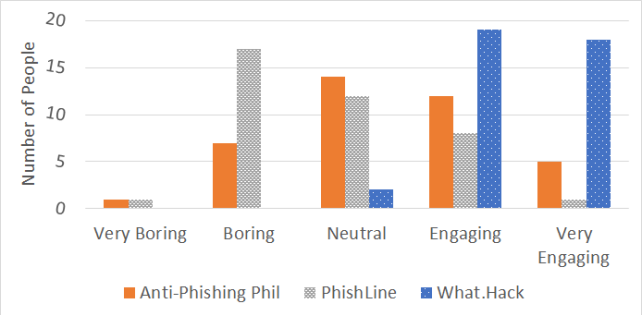

Figure 6: The engagement ratings. 95% of the participants find *What.Hack* being engaging or very engaging. The number is 44% for *Anti-Phishing Phil*, and 23% for *PhishLine*.

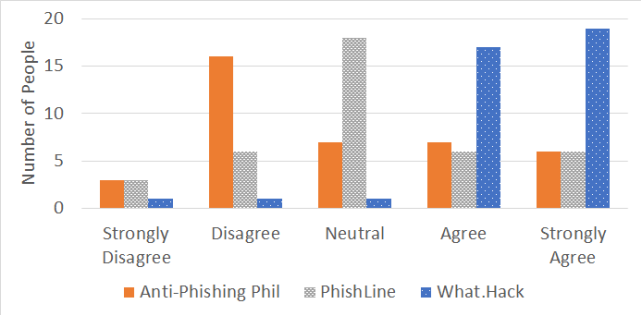

Figure 7: The recommendation ratings. 92% of the participants agree or strongly agree that they will recommend *What.Hack*. The number is 33% for *Anti-Phishing Phil*, and 33% for *PhishLine*.

that participants retained more conceptual knowledge by playing a game than watching videos. Unlike *What.Hack*, however, *Anti-Phishing Phil* players could not practice using the conceptual knowledge to vet phishing contents. While *What.Hack* could enhance anti-phishing self-efficacy in the user study, *Anti-Phishing Phil* failed to provide the same result. This feeling of "I know it, but I am unsure how to use it" could be the reason for the disparity in ratings for *Anti-Phishing Phil*. *What.Hack* avoided this issue, which implies that gamified cybersecurity education should teach both conceptual knowledge and how to use it.

Another implication is that gamified cybersecurity education should do the best to simulate all factors that affect learning outcomes to avoid unwanted side-effects. In our case, the cognitive load impacts susceptibility to phishing [66]. *What.Hack* prepares players for a heavy cognitive load that they would experience in handling real threats, but *Anti-Phishing Phil* does not. The *Anti-Phishing Phil* paper [56] reports that some players did not look further to avoid phishing attacks if the URL text did not raise suspicion, which did not happen to *What.Hack* players in the user study.

Limitations and future work. What.Hack did not fully utilize social engagement in situated learning like other role-playing games did for language learning [26] and science education [35]. Phishing, especially social engineering attacks, is a crime of deceiving people through online social interactions. Therefore, a multiplayer extension of *What.Hack* could potentially enhance anti-phishing training.

What.Hack has a broader design goal than *Anti-Phishing Phil* and *PhishLine*, which may introduce unfairness when comparing our game with these two conditions. We integrate conceptual knowledge into an email processing game and let the player internalize what kinds of URLs are dangerous, whereas the other two conditions only focus on conceptual knowledge, such as URL mechanics. However, we believe the most important consideration is the ability of anti-phishing training to prepare people to handle real email threats, and we found that our game is more effective in this regard. To further determine the impact of our game on long-term retention of phishing attack methods and defenses, we would like to conduct a field study in the future.

What.Hack offers a good starting point for the development of other similar game-based experiences in the field of cybersecurity education. For example, fake news is a trending global problem and new game designs for training people to identify fake news are rare and ad hoc [4]. Our game design can be re-skinned to introduce this issue. We would like to explore in the future whether such training is also more effective by using our game design.

6 CONCLUSIONS

The main goals of *What.Hack* are to 1) teach players anti-phishing techniques by simulating some of the circumstances in which people routinely defend against phishing attacks, 2) engage players by offering them freedom to experiment and observe narrative consequences, and 3) deliver a set of learning content through a task progression that starts easy and gradually grows more complicated. We presented results from a lab study demonstrating that *What.Hack* was able to improve players' correctness in identifying incoming threats by 36.7%, whereas a control group that played a different game did not achieve a statistically significant improvement. The results indicated that situated learning plays an important role in improving learning outcomes by engaging learners in a relatable simulation world.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers, Jo Iacovides, Amy Perelberg, Yuhang Zhao, Dongfang Gaozhao and Jiajing Guo for their valuable inputs and suggestions.

REFERENCES

[1] 2007. The Carnegie Cyber Academy - An Online Safety site and Games for Kids. http://www.carnegiecyberacademy.com/

[2] 2008. MAVI interactive. Agent Surefire. http://maviinteractive.com/mavi_products.asp. Accessed: 2018-09-20.

[3] 2016. Cyber Security Challenge UK Cyphinx. https://www.cybersecuritychallenge.org.uk/competitions/play-demand-cyphinx. Accessed: 2018-09-20.

[4] 2017. Game Sets Sights on Fake News. https://www.american.edu/soc/news/fake-news-game.cfm. Accessed: 2018-09-20.

[5] 2018. The Federal Bureau of Investigations, "Kids Games.". https://archives.fbi.gov/archives/fun-games/kids/kids-games. Accessed: 2018-09-20.

[6] 2018. Information Assurane Support Environment Cyber Protect. https://iatraining.disa.mil/eta/cyber-protect/launchcontent.html. Accessed: 2018-09-20.

[7] 2018. Information Security Office CMU "Anti-Phihsing Phil.". https://www.cmu.edu/iso/aware/phil/index.html. Accessed: 2018-09-20.

[8] 2018. OnGuardOnline. https://www.onguardonline.gov/media. Accessed: 2018-09-20.

[9] 2018. PhishLine Training. https://www.phishline.com/complimentary-content/

[10] Gupta BB Atawneh S. Meulenberg A. & Almomani E. Almomani, A. 2013. A survey of phishing email filtering techniques. In *IEEE communications surveys & tutorials*, Vol. 15.

[11] Erik Andersen, Eleanor O'Rourke, Yun-En Liu, Rich Snider, Jeff Lowdermilk, David Truong, Seth Cooper, and Zoran Popovic. 2012. The impact of tutorials on games of varying complexity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 59–68.

[12] John R Anderson, Lynne M Reder, and Herbert A Simon. 1996. Situated learning and education. *Educational researcher* 25, 4 (1996), 5–11.

[13] Nalin Asanka Gamagedara Arachchilage and Steve Love. 2013. A game design framework for avoiding phishing attacks. *Computers in Human Behavior* 29, 3 (2013), 706–714.

[14] Nalin Asanka Gamagedara Arachchilage, Steve Love, and Konstantin Beznosov. 2016. Phishing threat avoidance behaviour: An empirical investigation. *Computers in Human Behavior* 60 (2016).

[15] Ian Arawjo, Cheng-Yao Wang, Andrew C Myers, Erik Andersen, and François Guimbretière. 2017. Teaching Programming with Gamified Semantics. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

[16] Suranjith Ariyapperuma and Amina Minhas. [n. d.]. Internet security games as a pedagogic tool for teaching network security. In *35th Annual Frontiers in Education*. IEEE, S2D–1.

[17] Sasha Barab, Michael Thomas, Tyler Dodge, Robert Carteaux, and Hakan Tuzun. 2005. Making learning fun: Quest Atlantis, a game without guns. *Educational technology research and development* 53, 1 (2005), 86–107.

[18] Peter Chapman, Jonathan Burket, and David Brumley. 2014. PicoCTF: A Game-Based Computer Security Competition for High School Students.. In *3GSE*.

[19] Naomi C Chesler, Golnaz Arastoopour, Cynthia M D'Angelo, Elizabeth A Bagley, and David Williamson Shaffer. 2013. Design of a professional practice simulator for educating and motivating first-year engineering students. *Advances in Engineering Education* 3, 3 (2013), n3.

[20] Benjamin D Cone, Cynthia E Irvine, Michael F Thompson, and Thuy D Nguyen. 2007. A video game for cyber security training and awareness. *computers & security* 26, 1 (2007), 63–72.

[21] Dan Conway, Ronnie Taib, Mitch Harris, Kun Yu, Shlomo Berkovsky, and Fang Chen. 2017. A Qualitative Investigation of Bank Employee Experiences of Information Security and Phishing. In *Thirteenth Symposium on Usable Privacy and Security*.

[22] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, et al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–760.

[23] National Research Council et al. 2000. *How people learn: Brain, mind, experience, and school: Expanded edition*. National Academies Press.

[24] Mihaly Csikszentmihalyi. 1991. *Flow: The psychology of optimal experience*. Vol. 41. HarperPerennial New York.

[25] Gabriel Culbertson, Erik Andersen, Walker White, Daniel Zhang, and Malte Jung. [n. d.]. Crystallize: An Immersive, Collaborative Game for Second Language Learning. In *CSCW 2016*.

[26] Gabriel Culbertson, Shiyu Wang, Malte Jung, and Erik Andersen. 2016. Social Situational Language Learning through an Online 3D Game. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

[27] Andy Davis, Tim Leek, Michael Zhivich, Kyle Gwinnup, and William Leonard. 2014. The Fun and Future of CTF. In *USENIX Summit on Gaming, Games, and Gamification in Security Education (3GSE 14)*.

[28] Tamara Denning, Adam Lerner, Adam Shostack, and Tadayoshi Kohno. 2013. Control-Alt-Hack: the design and evaluation of a card game for computer security awareness and education. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*.

[29] Rachna Dhamija, J Doug Tygar, and Marti Hearst. 2006. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*.

[30] Julie S Downs, Mandy B Holbrook, and Lorrie Faith Cranor. 2006. Decision strategies and susceptibility to phishing. In *Proceedings of the second symposium on Usable privacy and security*.

[31] James Paul Gee. 2003. What video games have to teach us about learning and literacy. *Computers in Entertainment (CIE)* 1, 1 (2003).

[32] Mark Gondree and Zachary NJ Peterson. 2013. Valuing Security by Getting [d0x3d!] Experiences with a network security board game. (2013).

[33] Jason Hong. 2012. The state of phishing attacks. *Commun. ACM* 55, 1 (2012).

[34] Fares Kayali, Günter Wallner, Simone Kriglstein, Gerhild Bauer, Daniel Martinek, Helmut Hlavacs, Peter Purgathofer, and Rebbeca Wölfle. 2014. A case study of a learning game about the Internet. In *International Conference on Serious Games*. Springer, 47–58.

[35] Diane Jass Ketelhut, Brian C Nelson, Jody Clarke, and Chris Dede. 2010. A multi-user virtual environment for building and assessing higher order inquiry skills in science. *British Journal of Educational Technology* 41, 1 (2010), 56–68.

[36] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. 2009. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security*. ACM, 3.

[37] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*.

[38] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 905–914.

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

[40] Elmer Lastdrager, Inés Carvajal Gallardo, Pieter Hartel, and Marianne Junger. 2017. How Effective is Anti-Phishing Training for Children. In *Symposium on Usable Privacy and Security*.

[41] Jean Lave and Etienne Wenger. 1991. *Situated learning: Legitimate peripheral participation.* Cambridge university press.

[42] 3909 LLC Lucas P. 2013. Papers, Please: a dystopian document thriller. http://store.steampowered.com/app/239030/

[43] Jelena Mirkovic and Peter A. H. Peterson. 2014. Class Capture-the-Flag Exercises. In *USENIX Summit on Gaming, Games, and Gamification in Security Education (3GSE 14)*.

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

[45] Jason J Morrissette. 2017. Glory to Arstotzka: Morality, Rationality, and the Iron Cage of Bureaucracy in Papers, Please. *Game Studies* 17, 1 (2017).

[46] Casey O'Donnell. 2014. Getting played: Gamification, bullshit, and the rise of algorithmic surveillance. *Surveillance & Society* 12, 3 (2014), 349.

[47] Marc Olano, Alan T Sherman, Linda Oliva, Ryan Cox, Deborah Firestone, Oliver Kubik, Milind Patil, John Seymour, and Donna Thomas. [n. d.]. SecurityEmpire: Development and Evaluation of a Digital Game to Promote Cybersecurity Education.

[48] Cas Pars. 2017. *PHREE of Phish: The Effect of Anti-Phishing Training on the Ability of Users to Identify Phishing Emails.* Master's thesis. University of Twente.

[49] PwC. 2017. Game of Threats – A cyber threat simulation. http://www.pwc.com/us/en/financial-services/cybersecurity-privacy/game-of-threats.html

[50] Prashanth Rajivan and Cleotilde Gonzalez. 2018. Creative Persuasion: A study on adversarial behaviors and strategies in phishing attacks. *Frontiers in psychology* 9 (2018), 135.

[51] C Reigeluth and R Stein. 1983. Elaboration theory. *Instructional-design theories and models: An overview of their current status* (1983), 335–381.

[52] Andrew Ruef, Michael Hicks, James Parker, Dave Levin, Michelle L Mazurek, and Piotr Mardziel. 2016. Build It, Break It, Fix It: Contesting Secure Development. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 690–703.

[53] Wombat Security. 2017. State of the Phish. http://usdatavault.com/library/Wombat%20State%20of%20the%20Phish%202017.pdf

[54] David W Shaffer. 2006. Epistemic frames for epistemic games. *Computers & education* 46, 3 (2006), 223–234.

[55] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. 2010. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[56] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. 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*.

[57] Adam Shostack. 2017. Security Games & Resources. https://adam.shostack.org/games.html

[58] Michael F. Thompson and Cynthia E. Irvine. 2014. CyberCIEGE Scenario Design and Implementation. In *USENIX Summit on Gaming, Games, and Gamification in Security Education (3GSE 14)*.

[59] Jin-Ning Tioh, Mani Mina, and Douglas W Jacobson. 2017. Cyber security training a survey of serious games in cyber security. In *Frontiers in Education Conference (FIE)*. IEEE, 1–5.

[60] Endel Tulving and Donald M Thomson. 1973. Encoding specificity and retrieval processes in episodic memory. *Psychological review* 80, 5 (1973).

[61] Lev Semenovich Vygotsky. 1980. *Mind in society: The development of higher psychological processes.* Harvard university press.

[62] Chad Walker. 2015. Cryptomancer: A Fantasy Role-Playing Game about Hacking. http://cryptorpg.com/

[63] Rick Wash and Molly M Cooper. 2018. Who Provides Phishing Training?: Facts, Stories, and People Like Me. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 492.

[64] Gregory B White, Dwayne Williams, and Keith Harrison. 2010. The CyberPatriot national high school cyber defense competition. *IEEE Security & Privacy* 5 (2010), 59–61.

[65] Wikipedia. 2017. Podesta emails. http://en.wikipedia.org/w/index.php?title=Podesta%20emails&oldid=759435543.

[66] Emma J Williams, Amy Beardmore, and Adam N Joinson. 2017. Individual differences in susceptibility to online influence: a theoretical review. *Computers in Human Behavior* 72 (2017), 412–421.

[67] Che-Ching Yang, Shian-Shyong Tseng, Tsung-Ju Lee, Jui-Feng Weng, and Kaiyuan Chen. 2012. Building an anti-phishing game to enhance network security literacy learning. In *2012 IEEE 12th International Conference on Advanced Learning Technologies*. IEEE, 121–123.