Northumbria Research Link

Citation: Nicholson, James, Javed, Yousra, Dixon, Matt, Coventry, Lynne, Dele-Ajayi, Opeyemi and Anderson, Philip (2020) Investigating Teenagers' Ability to Detect Phishing Messages. In: 2020 IEEE 5th European Symposium on Security and Privacy Workshops (EuroS&PW). IEEE, Piscataway, NJ, pp. 140-149. ISBN 9781728185989, 9781728185972

Published by: IEEE

URL: https://doi.org/10.1109/eurospw51379.2020.00027 <https://doi.org/10.1109/eurospw51379.2020.00027>

This version was downloaded from Northumbria Research Link: http://nrl.northumbria.ac.uk/id/eprint/43530/

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: http://nrl.northumbria.ac.uk/policies.html

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher's website (a subscription may be required.)

Northumbria University NEWCASTLE

University Library

Investigating Teenagers' Ability to Detect Phishing Messages

line 1: 1st Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

line 1: 2nd Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

line 1: 3rd Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

***Abstract*—Young people are increasingly becoming responsible for the security of their devices, yet do not appear to receive formal instruction on how to protect themselves online. In this paper, we investigate the phish detection performance of teenagers while exploring how their familiarity with a service affects their overall performance in identifying phishing messages. Our study with 83 teenagers finds that teenagers were poor at distinguishing between genuine and phishing messages in an experimental task, yet performance was not affected by the familiarity of the service. However, our participants exhibited riskier behavior when making decisions on unfamiliar messages, suggesting that this is an area for further exploration. We discuss the implications of the poor phishing performance for teenagers and explore possible avenues to improve their awareness of these attacks, e.g. through embedding training content within the school curriculum.**

Keywords—social engineering, teenagers, phishing, cybersecurity, social factors

I. Introduction

With the advancement and ubiquity of technology, children in the Western world are being exposed to smartphones, tablets, and gaming systems from as early as 1 year old. By age 4, many children own a mobile device [1]. Similarly, children aged 5-15 years go online for at least 8 hours a week. Examples of common online activities for this age group include communicating through social media, playing games, and watching YouTube videos [2].

Increased exposure to the internet has highlighted the need for education directed to this age group, however, formal computing education is primarily geared towards personal safety (or *e-safety*) highlighting the issues of cyber predators using social media and games, sexting, searching and cyberbullying. Meanwhile, formal education for this age group about maintaining the digital security of accounts and devices (e.g. password management) is given far less priority and children are expected to protect their accounts and devices without any explicit cybersecurity knowledge [3].

While e-safety is of utmost importance for children, a more rounded cybersecurity education should be a growing priority as children are more frequently accessing their own devices and accounts on various services without adult supervision [1]. Even when adult supervision and guidance is provided, we know that many adults do not understand or follow cybersecurity best practices [4] which can result in passing on flawed practices.

One of the primary digital risks that children need to be aware of is phishing, a common social engineering attack ranked as one of the most dangerous online risks for children [5]. Phishing is an attempt to obtain sensitive information such as usernames, passwords, and credit card details by imitating a well-known service provider in an electronic communication such as an email. Children are a common target of phishing emails for the purpose of identity theft or the illegal use of children's personal information to steal money or credit [6]. In 2017, more than 1 million children (below 17 years of age) in the U.S. alone were victims of identity theft with estimated costs of $2.6 billion [7].

Several efforts have been made towards designing mechanisms and training tools to help protect people against phishing [8]–[11]. These tools have resulted in modest improvements in detection of phishing content, even if individuals and organizations continue to suffer from these attacks [12]. However, the effectiveness of these techniques in phishing detection performance has mostly focused on adults and university students, leaving young people under 18 as an understudied population.

Recently, Lastdrager et al. [13] took a first step towards using anti-phishing training for schoolchildren aged 8-13 years. The training improved children's short-term ability to distinguish phishing emails/websites from genuine emails/websites correctly, but this ability deteriorated over a period of 2-4 weeks. Similarly, Maqsood et al. [14] created and evaluated the effectiveness of a persuasive digital literacy game for children aged 11-13 years. However, to the best of our knowledge, phishing identification performance of teenagers aged 14-17 years has not been explored yet.

Teenagers form a particularly vulnerable group for a number of reasons. First, this is a developmental period where the brain is evolving and, as a result, a time of heightened vulnerability to risk taking behaviors in the pursuit of reward [15]. Less than ideal decisions and actions give rise to an increased incidence of unintentional injuries and violence, alcohol and drug abuse, unintended pregnancy and sexually transmitted diseases [16]. At a time of increased risk taking, teenagers start acquiring online accounts linked with payment methods, e.g. children's bank accounts are available from age 11 upward, while independent young person's accounts are available from age 16.

Secondly, teenagers' online footprints on social networks can be a valuable target for attack. As active users of social media websites, teenagers can often share a lot of data while interacting and communicating with each other. This sharing can be risky [17]. Additionally, by using publicly-available information obtained through social media, an attacker could potentially gain access to the young victim's network of family and friends, and invariably target them as well [13].

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Thirdly, teenagers have been identified as reliant on reactive thought processes more than reasoned processes. This means they may make decisions quickly, these may not be the best decisions and do not consider the potential for their decision being wrong. This type of thinking may make them more vulnerable to phishing emails as they may respond without considering that they are not genuine, and may be more reactive to the persuasive influences of urgency and panic [18], [19] , which many phishing emails attempt to instill [20].

Finally, it is not uncommon for older adults to rely on teenagers for setting up digital devices or for advice [21]. This reliance on younger people (usually teenagers who are perceived to be 'good' with technology) could conceivably result in compromises if teenagers are unable to correctly identify fraudulent messages.

While previous work shows that younger children are vulnerable to phishing [13], and that younger adults are also particularly vulnerable [9], it is important to understand whether these reported susceptibilities are due to these age groups having less experience with tested services. People's susceptibility to phishing is very contextual [22], hence it is important to consider the impact of service congruency. For example, will a phishing email from a familiar service be easier to identify because the individual is aware of what genuine messages look like, or will it be more difficult due to the individual having received regular messages from that provider? To this end, we investigate congruency as a factor in this study, splitting services into those that teenagers are likely to be familiar with and those that they are unlikely to encounter.

In this paper, we take a first step towards answering the following research questions:

RQ1: How accurately can teenagers detect phishing messages?

RQ2: How does the role of service congruency affect the accuracy of teenagers' phish detection?

RQ3: What is the relationship between teenagers' confidence in their ability to detect phishing messages and their actual performance?

Our first contribution is investigating the phishing susceptibility of a population that, to the best of our knowledge, has not been studied previously. As mentioned before, teenagers are of great interest as young people in this age group are actively participating in online social networks, beginning to take responsibility for financial accounts, and also provide advice to adults. We find that overall performance in the phish identification task was poor, despite our sample of teens having been primed about the experiment.

Our second contribution examines a new approach for conducting phishing experiments to better understand contextual factors. To the best of our knowledge, the role of message congruency in phishing has not been explored. We did not find a statistically-significant effect of congruency on teens' ability to detect phishing emails, but we did observe a response bias where our participants were more likely to err on the side of caution (e.g. identify as a phishing message) when evaluating congruent messages.

The rest of the paper is organized as follows. First, we give a brief background on phishing, the existing work on phishing prevention and existing efforts towards children security. We then describe our study design and methodology and present our results. Next, we discuss our findings, i.e., what we learnt from our study and the implications/future directions of this research. Finally, we conclude the paper with the main takeaways.

II. Related Work

A. Phishing

The number of phishing attacks has continually increased over the years. A recent survey [12] on nearly 15,000 end-users from 7 countries showed that 83% of the respondents had experienced a phishing attack in 2018 compared to 76% in 2017. Phishing severely impacts businesses; mid-sized companies pay an average of $1.6 million to recover from a successful phishing attack where the consequences include malware infections, compromised accounts, and data loss [12].

Phishing not only targets adults but also children and is listed as one of the top 7 digital risks for children and teenagers [5]. Children are a common target of phishing emails for the purpose of identity theft or the illegal use of children's personal information to steal money or credit [6]. More than 1 million children in the U.S. under the age of 17 fell victim to identity theft in 2017 costing approximately $2.6 billion [7].

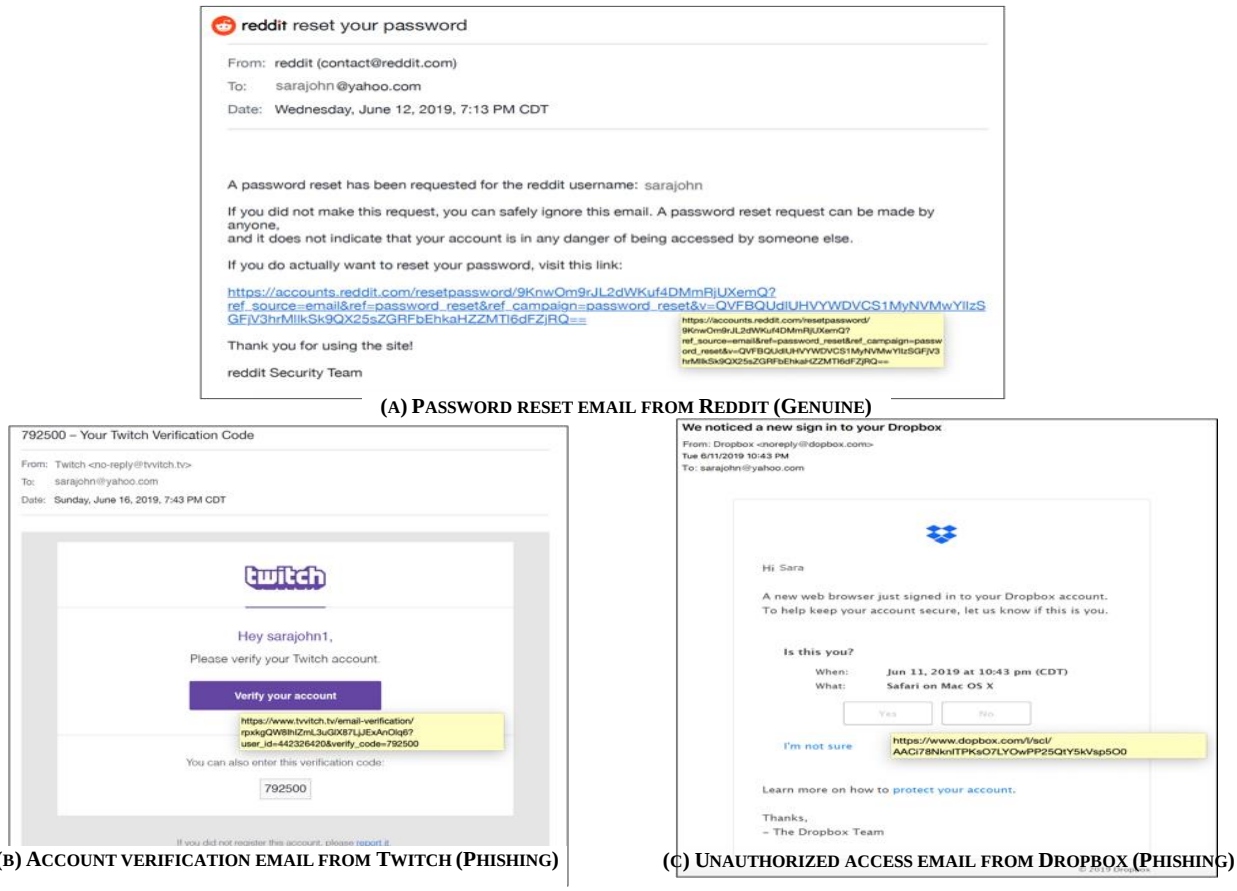
One of the reasons behind susceptibility to phishing attacks is poor cybersecurity knowledge and bad habits. Cain et al. [23] investigated the cyber hygiene of people aged 18 to 55 years and observed that younger adults have poor cybersecurity habits (or *hygiene*) related to password management and phishing. Adults have been shown to have poor calibration between confidence and actual performance when it comes to identifying phishing messages [11] which can then increase the likelihood that the attack is successful.

B. Anti-Phishing Efforts

With the prevalence and potential consequences of phishing, continuous efforts are made to improve cybersecurity knowledge of citizens and to develop protections against phishing attacks. Researchers have explored technical solutions, awareness through cybersecurity educational games and training materials, and the addition of cues in user interface to aid in phish detection.

Technical solutions have generally focused on systems which can identify suspicious websites, for example using browser plugins or identifying characteristic elements of a phishing email, e.g. [24]. Filtering algorithms can also bring improvements, e.g., [25]–[27], however, such phishing tools are not always accurate – some phish are missed and some genuine items are flagged as phish, i.e. there are problems with false positives and negatives [28] which either expose users to phishing emails or misdirect genuine emails to their spam folders.

Kumaraguru et al. [8], [9] designed two games, namely, Anti-Phishing Phil and PhishGuru to train users for detecting phishing websites. Participants who played the game were better able to identify fraudulent web URLs compared to the participants in the control group [8], [9], [29]. They also observed that younger adults aged (18-25) were more vulnerable to phishing attacks than older participants. Nyeste et al. [10] created anti-phishing training in a simple comic and



(A) PASSWORD RESET EMAIL FROM REDDIT (GENUINE)

(B) ACCOUNT VERIFICATION EMAIL FROM TWITCH (PHISHING)

(C) UNAUTHORIZED ACCESS EMAIL FROM DROPBOX (PHISHING)

FIGURE I. EXAMPLES OF GENUINE AND PHISHING EMAILS FROM CONGRUENT AND NON-CONGRUENT SERVICES USED IN OUR STUDY DATASET (YELLOW SQUARES SHOW THE EMBEDDED LINKS THAT APPEAR WHEN HOVERING OVER HYPERLINKS)

complex video game form and showed that both were helpful in decreasing phishing susceptibility as measured by miss rates for college-aged and computer-savvy participants.

More subtle approaches have also been explored. For example, Nicholson et al. [11] investigated the use of "nudges" – or encouragements – to help with phish identification. Their study on adults showed that highlighting relevant information (salience) proved to be effective in aiding the user's accuracy in phish identification. Miyamoto et al. [30] developed an eye-tracking based system to inculcate the habit of looking at the URL address bar before entering sensitive information in the website's input fields. The system first de-activates the input fields in a website and using eye-tracking data determines if the user has looked at the website's URL. The input fields are activated after confirming user attention on the URL address bar. Their system showed good learnability and improved the accuracy of detecting phishing websites when tested on university students.

C. Cybersecurity, Children, and Teenagers

Over the past few years, researchers have started paying attention to digital risks for children and educating them about cybersecurity. Existing studies in this domain have explored children's password management strategies, the design of anti-phishing training, and web-based games for improving children's digital literacy.

Children use technology and online services that require frequent authentication. However, currently there is sparse research on how children authenticate and protect themselves online. Choong et al. [31] surveyed password knowledge and practices of children aged 8-15 years, observing that children showed confusion between the concepts of passwords, privacy and safety or protection. Age influenced children's password practices, with the younger children relying on parents for creating and remembering passwords while older children created longer passwords compared to younger children. Maqsood et al. [32] studied how children aged 11-13 years created passwords given different password policies. They observed that children create simple passwords containing their personal information such as name or age and believe that their passwords would be hard for a stranger to guess. Similarly, other researchers have observed that children create passwords containing whole words and personal information and have trouble recalling long and complex passwords compared to simple ones [33], [34]. Another work [35] has observed that 6-12 year olds are not necessarily better with graphical passwords – alternatives to passwords which have traditionally been considered easier to use for certain populations [36]. In fact, surprisingly, their study on 13 participants showed that the success rate for the click-based graphical password was lower than that for textual passwords. These studies collectively demonstrate that both younger and older children demonstrate poor cybersecurity hygiene and suggest that there may be issues with how this information is being taught to this age group.

Children need education on digital risks appropriate for their age. Lastdrager et al. [13] investigated the effectiveness of anti-phishing training for school children aged 9-12 years. The

training was done in a story-telling format during lecture time. The children were then given a paper-based test to classify a set of emails and websites as either phishing or genuine. The training proved to be effective in increasing phishing identification. However, this performance decayed after a period of 2 and 4 weeks. Their work highlights the need for regular anti-phishing training tailored towards children at different levels of education. More generally, Maqsood et al. [14] created a web-based game to improve 11-13-year-old children's digital literacy on cyberbullying, tracking, privacy, sharing, and authentication of information. The participants demonstrated the usefulness of the game and even retained the information after a period of 1 week. This collection of studies identifies that children can be taught cybersecurity topics effectively given appropriate training for their age which could be integrated into mainstream education via IT classes [37]. While some of the recent work discussed above has begun to study children aged 9-13 years, teenagers still remain unstudied when it comes to phishing. The only exception relates to work exploring methods for improving phishing materials for teens, but no phish detection is performed by the participants [38], [39]. Therefore, we include children aged 13-17 years in our sample.

III. Method

Our study set out to assess how teenagers performed on a simple phishing test while considering contextual indicators by manipulating the congruency of the service.

The study consisted of a repeated measures design where participants were shown both congruent and non-congruent emails in mixed and randomized order. The variables under investigation were the accuracy of the choice (genuine or fake) and the confidence score (0-100%) for each email.

Congruent emails were classed as those that our participants had a high likelihood of encountering on a regular basis. Based on demographic information reported by marketing portals online and real-world observations of young people's technology use (e.g. [40], [41]) we decided on the following six services as congruent for this population: Snapchat, Instagram, YouTube, Twitch.tv, Spotify, and Reddit.

Non-congruent emails were classed as those that our participants had a low likelihood of encountering on a regular basis. Based on demographics reported by marketing portals online and real-world observations [40], we decided on the following six services as non-congruent: Amazon, eBay, PayPal, Dropbox, Twitter, and Netflix.

In order to verify the congruency of messages for our participants, we asked them to complete a three-point scale for each message consisting of "I don't use this service", "I use this service sometimes", and "I use this service all the time". Non-congruent services were defined as those that they do not use at all, while congruent services comprised the other two categories given that very few participants used a service only "sometimes". We note that based on this self-reported familiarity two services were swapped for the analysis: Amazon was categorized as congruent and Reddit as non-congruent.

A. Materials: Development of Email Messages

Twelve email messages depicting twelve distinct web services (described below and in Table I) were developed in the format of interactive PDFs where participants could hover over links to reveal embedded web links (see Figure I). The study itself was hosted on Qualtrics[1] to ensure that all messages were counterbalanced across participants and facilitated the collection of questionnaire data following the experiment.

All email messages were based on existing genuine messages[2]. Genuine messages were kept as standard with minor changes to ensure consistency (i.e. changing the name of the receiver). The phishing emails had the sender name, email, and links changed to reflect existing phishing emails (based on the popular PhishTank[3] database) – with minor alterations for safety (i.e. slightly altering the web link to prevent unintended website visits).

The phishing messages were constructed to be simple to identify as long as appropriate techniques were employed: checking the sender email address and checking the destination of web links. We deployed an equal distribution between links displayed in cleartext and those masked behind other text or images. Then, we used four common techniques [42] for transforming the genuine links into phishing links: masking the domain by using subdomains (e.g. *paypal.fakesite.com*), using common character substitutions (e.g. *paypai.com*), dropping characters (e.g. *paypa.com*), and using incorrect top-level domains (e.g. *paypal.co*).

TABLE I
EMAIL MESSAGES USED IN EXPERIMENT

Service	Classification	Type
Congruent		
Snapchat	Genuine	Verify
Instagram	Genuine	Unauthorized
YouTube	Phishing	Unauthorized
Twitch.tv	Phishing	Verify
Spotify	Phishing	Reset
Amazon	Genuine	Reset
Non-Congruent		
Reddit	Genuine	Reset
eBay	Genuine	Verify
PayPal	Phishing	Verify
Dropbox	Phishing	Unauthorized
Twitter	Genuine	Reset
Netflix	Phishing	Unauthorized

We chose three types of action messages based on common phishing email tactics: password reset, unauthorized access, and email verification (see Figure I for examples). Password reset emails presented users with a (requested) link to reset their current password for the service. Unauthorized access emails notified the user that a sign-in had been observed in a foreign location and gave the user a link to follow if the login

[1] www.qualtrics.com
[2] Messages available at: [redacted for review]
[3] www.phishtank.com

was not recognizable. Finally, the account verification email presented the user with a link to verify their sign up with a service (e.g. after initially signing up).

B. Participants

We recruited 83 teenagers aged 12-17 through a university outreach program in association with regional law enforcement in the UK. Initially, an advert for the overall university event focusing on cybersecurity (consisting of various presentations, demonstrations, and experiments) was distributed across schools in the region. Any interested schools were instructed to contact the university to secure a place in the event by detailing the number of students to attend. Once all places were filled, information about this specific study was sent to schools (see subsection D below for more information). This resulted in the attendance of teenagers from 14 schools from across the region. A post-hoc statistical power analysis [43] suggests an adequate sample size to detect a medium effect size given our simple experimental design.

We were unable to collect exact ages or gender due to ethical safeguarding, however, we asked participants to select an age range instead. Our sample resulted in 60 teenagers 12-14 years old, and 23 teenagers 15-17 years old.

C. Procedure

All data was collected as part of two group testing sessions, where approximately 40 teenagers completed the task individually gathered in one computer room. The task lasted approximately 15 minutes.

Upon their arrival, participants were briefed and given the web link to the experiment (see subsection D below). Participants viewed the 12 emails on the desktop computer in random order and were asked to select an action for each message (action or delete), along with their confidence in their decision (0-100%) and their familiarity with the service ("I don't use this service" to "I use this service all the time").

Once all participants completed the task, the instructor revisited each email on the main projection screen and highlighted the cues and techniques for correctly classifying them, specifically verifying the sender's email address [11, 44] and checking the destination of links. At this point, questions were also answered. Finally, all participants were debriefed before leaving the room.

D. Ethics

All our participants were under 18 years of age. Therefore, we sent out an information sheet to all parents (through the school) making them aware of the proposed study. Parents were given two weeks to withdraw their children from the study if they wished by contacting their schoolteacher.

A teacher from each school attended the testing session (a total of 14). Teachers were allowed to complete the phish detection task, but their answers were not analyzed.

Before the start of the study, all participants were given the option to take part. If they agreed, they were given the web link to access the study. Once the online study was loaded, participants once again had to agree to take part in the study. It was made clear that they could withdraw at any point by simply closing the browser tab. Due to the lack of deception around the task, all participants were primed to scrutinize messages.

Following the testing session, all participants were given the correct answers and told how to evaluate emails for authenticity. This was an important step to ensure that teenagers could increase their understanding of phishing and improve their behaviors – and their protection – in the future. This debrief was followed by a talk by law enforcement reiterating the dangers of the internet and the consequences of participating in online criminal activities. This study was approved by the University's Ethics Committee.

IV. Results

Considering that participants were fully aware of the purpose of the experiment, the overall success rate (*correct/total*) for our young participants was poor. The success rate for detecting phishing emails (*correct phishing/total phishing*) was 70%, while, the success rate for detecting genuine emails (*correct genuine/total genuine*) was 50%. The success rate for congruent messages was 57.9% while the success rate for non-congruent messages was 59.5%.

A. Scoring

We used signal detection theory to analyze the performance differences between congruent and non-congruent messages. Signal detection theory was originally developed to determine the sensitivity of a participant to the presence of a target (phishing emails) against a background of noise (genuine emails). Simple measures, such as success rate, ignore false positives which have become increasingly problematic for both individuals and organizations. As such, the user judgement of genuine/phish messages was scored in terms of classical signal detection theory, i.e. as a hit, a miss, a true negative or a false positive. In our task, a phish identification experiment, we categorize phishing emails as the signal (the desired selection) and genuine emails as noise, thus: *Hit* refers to phishing emails that were correctly identified as phishing emails; *False Positive* (or false alarm) refers to genuine emails that were incorrectly identified as phishing; *Miss* refers to phishing emails that were identified as genuine emails; *True Rejection* refers to genuine emails that were identified as such by the participant.

The discriminability index *d'* is a statistic used in signal detection that provides the separation between the means of the signal and the noise distributions in units of standard deviation of noise distributions. The response bias, criterion (*c*), reflects how biased users are towards treating a stimulus as signal or noise. It is measured by how far their decision threshold is from the intersection of the two distributions. A negative response bias ($c < 0$) reflects a tendency to call uncertain stimuli signals. With phishing as the signal, negative values of *c* reflect a tendency to call uncertain messages phishing, indicating greater aversion to misses (treating phishing messages as genuine) than to false alarms (treating genuine messages as phishing). We refer readers to relevant texts (e.g. [45]) for further information on this method, and to [11], [46] for examples of signal detection being used in phish detection.

B. Sensitivity and Response Bias

We conducted a two-tailed t-test with participants' *d'* scores (see Table II) to understand whether the familiarity with the service had any effect on their ability to distinguish the

trustworthiness of messages. We found no significant statistical difference in young people's phishing detection capabilities (*d'*) between congruent and non-congruent email messages, $t(82)=0.494$, $p=.622$. This suggests that seeing email messages more often, or having more awareness of that service, may not affect a young person's ability to successfully evaluate its validity.

TABLE II
SENSITIVITY (*d'*) DESCRIPTIVES (HIGHER IS BETTER)

	Mean	N	Std. Dev.
Congruent messages	.60	83	1.45
Non-Congruent messages	.70	83	1.60

We then conducted a two-tailed t-test with participants' *c* scores (see Table III) to understand whether their decision-making approach differed depending on the familiarity of the service. We found a significant statistical difference in teenagers' approaches to decision making, where participants were more 'liberal' when presented with congruent emails, $t(82)=2.248$, $p=.027$. Simply, teenagers were more likely to err on the side of risk when it came to non-congruent emails and on the side of caution when identifying congruent emails. However, the general trend for both congruent and non-congruent email classifications was on a conservative range (i.e. greater than zero).

TABLE III
DECISION BIAS (*c*) DESCRIPTIVES (0< = CONSERVATIVE, 0> =LIBERAL)

	Mean	N	Std. Dev.
Congruent messages	.16	83	0.95
Non-Congruent messages	.41	83	0.84

C. Confidence

We noted earlier the importance of well-calibrated confidence in making risk decisions. In this study, we measured user confidence in each email judgment and map these with actual performance as can be seen in Figures 2 & 3.

We then constructed confidence calibration curves for both the phishing and the genuine emails. A calibration curve is a graph where subjective confidence of being correct is plotted against the actual performance (in this case confidence percentage is measured against accuracy percentage). The curves are created by computing the mean accuracy of those items where participants have given a particular confidence score. On each figure, the diagonal or *identity line* shows perfect calibration. Any data points above this line show under-confidence and points below the line show over-confidence. To take one example, a data point that shows 80% on the x-axis and 40% on the y-axis is showing that when we aggregate those emails in which the mean confidence rating is 80%, the mean accuracy rate for those same emails is only 40% (i.e. participants are over-confident). Thus, good calibration would be indicated by data curves forming close to the diagonal *identity line* and poor calibration would be shown by a deviation from this line. For more information, on calibration and phishing emails, see [11].

If we look first at the calibration curves for genuine emails (Figure II) then we can see that over-confidence predominates – participants are generally less accurate than they believe

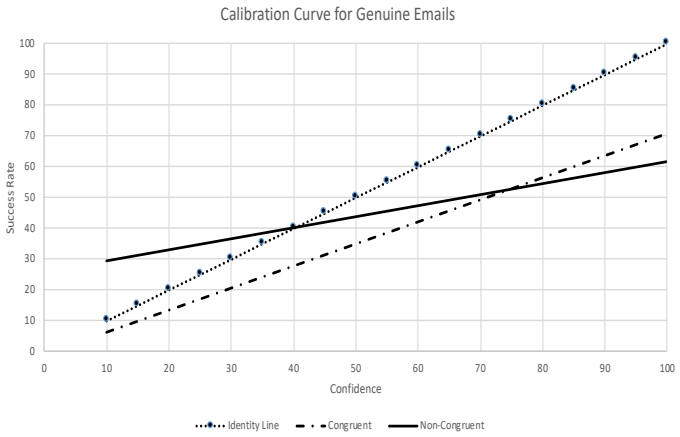


FIGURE II. CALIBRATION CURVE FOR GENUINE EMAILS. THE IDENTITY LINE SHOWS PERFECT CALIBRATION WITH UNDERCONFIDENT RESPONSES PLOTTED ABOVE AND OVERCONFIDENT RESPONSES PLOTTED BELOW

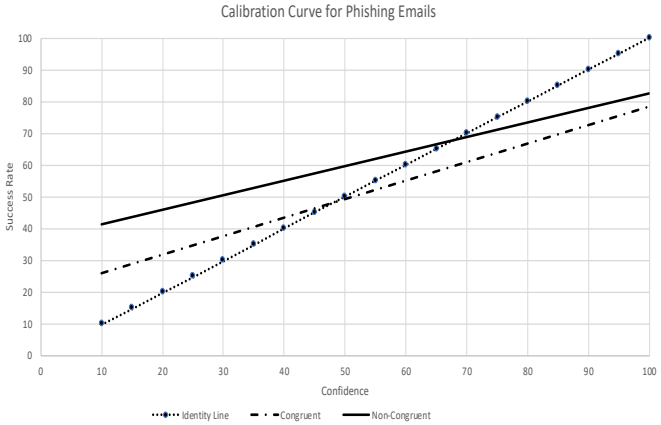


FIGURE III. CALIBRATION CURVE FOR PHISHING EMAILS. THE IDENTITY LINE SHOWS PERFECT CALIBRATION WITH UNDERCONFIDENT RESPONSES PLOTTED ABOVE AND OVERCONFIDENT RESPONSES PLOTTED BELOW

themselves to be. Turning now to the calibration curves for phishing emails (Figure III), we can see how surprisingly well-calibrated user confidence is for these items– i.e. the lines resemble that of the identity line. This is perhaps supported by the fact that our participants performed better when identifying phishing emails than when evaluating genuine emails.

D. Age Comparison

Here we present a brief comparison of the performance between younger and older teenagers in order to understand whether age could have influenced their ability to distinguish between phishing and genuine emails. However, given the uneven numbers of participants across age groups (12-14 = 60; 15-17=23), we only report descriptive statistics and cannot make any evidence-based claims other than to suggest further investigations.

There does not appear to be a noticeable difference between the performance of both age groups (see Table IV), which is surprising to an extent. Older teenagers (15-17) had an average success rate of 59.5%, with the success rate for phishing emails specifically at 69% and genuine emails at just 50%. Younger teenagers (12-14) had nearly identical metrics, with an overall success rate of 59% and success rates for phishing and genuine emails at 68% and 50% respectively. When looking at the performance of both groups with congruent and non-congruent

services we see some slight differences, but we refer readers to Table IV for full breakdowns.

TABLE IV
PHISH DETECTION SUCCESS RATE MEANS ACROSS DIFFERENT AGE GROUPS

Age	n	Overall	Phish	Genuine
12-14 years old	60	59%	68%	50%
15-17 years old	23	59.5%	69%	50%
Congruent Messages				
12-14 years old	60		64%	52%
15-17 years old	23		69%	51%
Non-Congruent Messages				
12-14 years old	60		73%	48%
15-17 years old	23		69%	49%

E. Types of Email Messages

In order to understand the role of message type on phish detection accuracy, we conducted a CHI Square test. We observed a statistically significant association between email type and phish detection accuracy, $\chi 2(2) = 22.212$, $p < .001$, where participants were more accurate with messages focused on unauthorized account access as compared to password reset and account verification. This is encouraging, as it suggests that teenagers are able to differentiate between an actual attack on their account and a social engineering attempt, with the caveat that we do not know whether participants would have acted on the genuine messages in the wild and investigated the service or simply ignored it.

TABLE V
PHISH DETECTION SUCCESS RATES ACROSS DIFFERENT EMAIL MESSAGES

Message Type	Mean
Unauthorized Access	69.8%
Email Verification	54.8%
Password Reset	53.3%

However, the fact that performance with password reset messages was so poor is alarming. Notably, our participants performed particularly poorly when identifying the genuine password reset emails, which suggest that they associate these types of emails with phishing. That can be positive – but presents the danger that if their account is genuinely under attack then they would not recognize it. As mentioned previously, ignoring messages could also be problematic.

When comparing the performance of participants with individual emails, we see a relative comparable performance across most services. The single exception is the Netflix email, where participants scored a high 82% success rate – although it is unclear why our participants were more accurate with this message. For comparison purposes, the YouTube email (also phishing and also unauthorized access, but non-congruent) had a 69% success rate.

V. DISCUSSION

This paper set out to understand how teenagers performed on a simple phishing test. Specifically, we looked to answer three research questions:

RQ1: How accurately can teenagers detect phishing messages?

Overall phishing detection performance was poor, with an overall success rate of 59% and sensitivity scores (*d'*) of .60 and .70 for congruent and non-congruent messages respectively. It is important to remember that misclassifying only a single phishing message can lead to irreversible consequences such as identity theft, and therefore a success score of less than 2/3 is worrying. This poor performance could be a result of the risk-taking traits identified within this population [15], [16], [19], a lack of awareness of how to identify phishing messages [13], or possibly a combination. However, to put some of these figures in some context, the sensitivity scores reported here are not too dissimilar from those reported by another phishing study that used signal detection theory: Nicholson et al. [11] reported a *d'* of .60 for adults, albeit using different stimuli, age groups, and genuine/phishing ratios.

We briefly compared the mean performance between two age ranges, 12-14 and 15-17 years old, and found near-identical performances – with the caveat that these results may not be representative due to uneven samples. Regardless, this apparent lack of improvement with age is worrying. It is not unreasonable to expect children to learn more about how to identify fake messages either through exposure or instruction over time, at school or at home. However, this lack of improvement could signal that schools' current approaches to prioritizing e-safety over cybersecurity training (e.g. [31]) may be leaving young people vulnerable to a different category of risks. Indeed, previous work has reported that younger adults have the worst cybersecurity behaviors amongst all age groups [23], and it is reasonable to argue that this stems from a lack of instruction at a formative age.

Our findings also suggest that teens may ignore some messages when they are unsure. This is typically seen as positive behavior, emphasizing caution over risk. However, conservative behavior can also be problematic when it is due to ignorance – for example, if an account is under genuine attack, and the user is unable to remedy this attack by, e.g. changing their password, then it may lead to an eventual compromise. Applied to a work context, such behavior could result in employees ignoring genuine emails and costing their organization money in lost business and/or fines.

RQ2: How does the role of service congruency affect the accuracy of teenagers' phish detection?

We found that teenagers' performance of email classification did not differ between messages that were derived from congruent services (those they are likely to use) and non-congruent services (those that they were unlikely to use). While this may appear as a surprising result – after all, it seems logical that people should be able to spot mistakes in content that they are familiar with – upon further scrutiny it can be explained by the fact that overall detection was very poor. In essence, it appears that participants were unable to spot the key verification cues (i.e. sender email address and link authenticity) and thus the congruency of the message became irrelevant. A key exception here was Netflix, where

participants had an 82% success rate, the highest amongst all services by a margin of 10%.

We found that message congruency played a role in how teenagers approached the task. Participants had a riskier strategy when faced with non-congruent messages – that is, they were more likely to classify an email as 'genuine'. On the other hand, teenagers were more likely to classify congruent emails as 'phish', demonstrating a more conservative strategy. However, it is important to note that the overall selection trend favored a risky approach. This insight helps explain why participants generally performed better at detecting phishing messages over genuine ones.

The implications for the role of congruency in phish detection are not clear. On the one hand, there was no statistically significant difference in the sensitivity of teenagers to congruent or non-congruent messages, but on the other hand they did appear to favor a riskier approach with unfamiliar services. This suggests that the paradigm may be worth exploring further to understand whether the gradient of congruency could influence the decision making, for example looking at very familiar vs. less familiar messages rather than familiar vs. unfamiliar as we have done.

RQ3: What is the relationship between teenagers' confidence in their ability to detect phishing messages and their actual performance?

Participants demonstrated over-confidence when labelling genuine emails i.e., they were less accurate than they believed themselves to be. However, their confidence was well-calibrated when labelling phishing emails. In reality, when combined with their overall performance on the task, these confidence measures seem to suggest that our participants were unsure what cues to look out for and chose 'phish' due to being primed by the context of the experiment (a phish identification task). Sadly, this indicates that under non-experimental settings their performance is likely to be worse.

A. Towards Improving Teenagers' Phishing Detection Capabilities

Our results suggest that teenagers need support for dealing with social engineering attacks. As established in the literature, age-appropriate training is a useful tool in educating children about phishing and how to identify phishing emails [13]. However, for maximum reach and impact, the training should to be integrated into mainstream education with the help of national and local policymakers.

A key aspect to integrating phishing training into mainstream education is to involve teachers in the training itself so they can feel confident and comfortable talking about these issues and countermeasures in their classrooms. Teachers, of course, are important role models in the education of young people: Their behaviors and opinions on technology are likely to be noticed by pupils and in some cases copied [47]. Additionally, for any positive change to be sustainable, teachers first need to be aware of the status quo, understand the need for change, and then be actively involved in working towards the change [48].

With this in mind, we set out to understand whether a simple training awareness session (i.e. observing the students perform the phishing exercise) could motivate teachers to take action. As a follow-up exercise to the phishing experiment, we asked the teachers who supervised the students to share their experiences of the event and the material. Three months following the initial session, we sent out an online questionnaire to teachers asking them about their observations since the event and their perceptions of any changes in their students' or their own knowledge and confidence with the subject material. The questionnaire consisted of 9 questions with a mix of 5-point Likert scale questions addressing confidence in the subject since the event, open-ended questions on their follow up discussions with management, and binary yes/no questions addressing their actions since the event. In total, we received responses from seven of the nine teachers who were contacted.

The responses from the teachers who attended the event were overwhelmingly positive. First, the teachers' responses suggest that observing the teenagers complete the phishing identification task and listening to the follow-up explanations from the researchers improved their self-efficacy and confidence: All of the surveyed teachers said they were confident or very confident in their ability to help young people identify and deal with scam email messages since the event. Likewise, all of the teachers said they felt their students were confident or very confident about identifying and dealing with scam email messages since the event. This supports other findings from literature that suggest that training can be effective in educating people about phishing attacks [13].

In the same vein, teachers were asked how they felt about adding content around phish identification techniques such as emphasizing the importance of looking at the sender's email address (e.g. [11]) to the classroom curriculum. More than 5/7 of the teachers said they *felt strongly* that this material should be taught in the everyday classroom. The results also show that in the following 3 months, 4/7 teachers had suggested to others in their school that similar content should be delivered as part of the curriculum, from colleagues to the IT department and all the way up to headteachers. Overall, this feedback confirms that the teachers saw value in the techniques we demonstrated to the students, not just as a one-off event, but as something that was useful and important enough to be integrated into the school curriculum.

Perhaps the most promising objective measure is the teachers' reported intention to deliver relevant material in the future: 6/7 of our responding teachers said they were planning to teach similar content in their classroom as a result of their experience. Despite only two of the teachers having actually taught similar contents to their students, given the short follow up timeframe, it is likely that appropriate opportunities may not have presented themselves organically during this time.

Another follow up with the teachers ten months following the study found that their views remained unchanged with regards to the importance of the subject being taught in school. However, three teachers had incorporated content on phishing into their teaching. We found that teachers in technical subjects (e.g. computer science, IT) had in fact improved on their practice, but those not teaching these subjects were still unsure how to do this. One teacher explained how as a result of the session she now uses demonstrations to teach phishing, e.g. by

going through real emails in her inbox and spam folder and pointing out the different techniques the attackers use in the actual emails, supporting recent work suggesting that demonstrations can be an effective method for cybersecurity training. Importantly, we find that teachers felt confident in incorporating good practice into their teaching: *"we don't go into a lot of depth about what to do after, as that's not part of the curriculum, but now we talk about taking a moment to make a determination about whether it's genuine or not"*.

Although we only report on the views from a small number of teachers, these are important insights into cybersecurity education within mainstream education, and seems to suggest that if school teachers are trained on the subject, they could then have the confidence to share with students as part of their academic learning journey. Of course, more work is needed to understand the long-term benefits of this approach, and to explore the best methods for supporting non-technical teachers in emphasizing good cybersecurity habits to students.

B. Limitations

Our study is not without limitations. This study suffers from issues similar to most other lab-based phishing experiments: the relevance and accuracy of the materials. In our case, all emails – both genuine and fake – were based on real emails sent by the relevant services and they closely resembled the source material. In this sense, our materials could be considered more sophisticated than most phishing and genuine emails seen in the wild, where typical 'subjective' cues of phishing such as spelling and grammar mistakes, poor formatting or general inconsistencies were not available to participants. However, our messages are unlikely to be as effective as spear phishing attacks as only simple manipulation on the sender address and the destination URLs were made.

Similar to other lab-based phishing experiments, our participants were aware that they were participating in a phishing test. This priming possibly influenced participant responses and made them more careful in evaluating the email messages – quite possibly manifesting in the high calibration between performance and confidence with the phishing emails, and overall better performance identifying the phishing messages over the genuine ones. However, despite this extra scrutiny, the overall performance was poor and raises concerns over the ability of teenagers to detect phishing emails under non-experimental conditions.

We recruited participants from 14 regional schools in England. Therefore, our findings may not generalize to all teenagers but does appear to be consistent with existing literature detailing poor phish detection ability by younger children [13] and young adults [9]. Additionally, the recruitment of the teenage participants was not under our control and relied on the schools themselves promoting the event and encouraging students to attend.

We were also unable to follow up with our participants, and this prevented us from discovering the true cause of their poor performance. While we can use service congruency and self-reported confidence to understand their selections in more detail, actual interviews with the teenagers may have yielded further insights. Future work could explore the reasons behind the poor performance of teenagers in phish detection tasks but perhaps also explore other methods for improving behaviors such as being conscious of what web addresses and services they interact with regularly, and knowing how to read these effectively.

Finally, the concept of service congruency is subjective, and this is part of what makes phishing such a challenging phenomenon to study. We controlled for this subjectivity by verifying with our participants during the task and in our case, our choice of services was largely appropriate. However, it may be difficult to scale up effectively with other more diverse populations, but utilizing fake services for the non-congruent condition could lead to a more scalable solution.

REFERENCES

[1] H. K. Kabali *et al.*, "Exposure and use of mobile media devices by young children," *Pediatrics*, vol. 136, no. 6, pp. 1044–1050, Dec. 2015.

[2] ofcom and Oftcom, "Children and parents: Media use and attitudes report," *Ofcom*, no. November, p. 228, 2015.

[3] J. Orlando, "Kids need to learn about cybersecurity, but teachers only have so much time in the day," 2019. [Online]. Available: https://theconversation.com/kids-need-to-learn-about-cybersecurity-but-teachers-only-have-so-much-time-in-the-day-112136. [Accessed: 29-Feb-2020].

[4] J. Nicholson, L. Coventry, and P. Briggs, "Introducing the Cybersurvival Task: Assessing and Addressing Staff Beliefs about Effective Cyber Protection", in *SOUPS 2017*.

[5] Kaspersky, "Internet Safety for Kids: Top 7 Internet Dangers," 2019. [Online]. Available: https://usa.kaspersky.com/resource-center/threats/top-seven-dangers-children-face-online. [Accessed: 29-Feb-2020].

[6] Department of Homeland Security, "Kids Cyber Security Presentation." [Online]. Available: https://www.dhs.gov/sites/default/files/publications/Kids Cybersecurity Presentation.pdf.

[7] K. Grant and CNBC, "Child identity theft is a growing and expensive problem," 2018. [Online]. Available: https://www.cnbc.com/2018/04/24/child-identity-theft-is-a-growing-and-expensive-problem.html. [Accessed: 29-Feb-2020].

[8] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong, "Teaching Johnny not to fall for phish," *ACM Trans. Internet Technol.*, vol. 10, no. 2, pp. 1–31, May 2010.

[9] P. Kumaraguru *et al.*, "School of Phish : A Real-World Evaluation of Anti-Phishing Training Categories and Subject Descriptors," *Symp. Usable Priv. Secur. - SOUPS*, p. 12, 2009.

[10] P. G. Nyeste and C. B. Mayhorn, "Training Users to Counteract Phishing," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 54, no. 23, pp. 1956–1960, 2010.

[11] J. Nicholson, L. Coventry, and P. Briggs, "Can we fight social engineering attacks by social means? Assessing social salience as a means to improve phish detection", in *SOUPS 2017*.

[12] Wombat Security Technologies, "State of the Phish," 2019.

[13] E. Lastdrager, I. C. Gallardo, M. Junger, and P. Hartel, "How Effective is Anti-Phishing Training for Children?", in *Conference on Human Factors in Computing Systems (CHI)* 2017.

[14] S. Maqsood, "Evaluation of a persuasive digital literacy game for children," in *Conference on Human Factors in Computing*

Systems - Proceedings, 2018, vol. 2018-April, pp. 1–6.

[15] M. Gardner and L. Steinberg, "Peer Influence on Risk Taking, Risk Preference, and Risky Decision Making in Adolescence and Adulthood: An Experimental Study," 2005.

[16] B. J. Casey, R. M. Jones, and T. A. Hare, "The adolescent brain," *Annals of the New York Academy of Sciences*, vol. 1124. Blackwell Publishing Inc., pp. 111–126, 2008.

[17] E. Christofides, A. Muise, and S. Desmarais, "Risky Disclosures on Facebook: The Effect of Having a Bad Experience on Online Behavior," *J. Adolesc. Res.*, vol. 27, no. 6, pp. 714–731, 2012.

[18] J. Arnett, "Reckless Behavior in Adolescence: A Developmental Perspective," 1992.

[19] D. B. Branley and J. Covey, "Risky behavior via social media: The role of reasoned and social reactive pathways," *Comput. Human Behav.*, vol. 78, pp. 183–191, Jan. 2018.

[20] A. Ferreira and G. Lenzini, "An Analysis of Social Engineering Principles in Effective Phishing," *Proc. - 5th Work. Socio-Technical Asp. Secur. Trust. STAST 2015*, pp. 9–16, 2015.

[21] J. Nicholson, L. Coventry, and P. Briggs, "'If It's Important It Will Be A Headline': Cybersecurity Information Seeking in Older Adults," in *Conference on Human Factors in Computing Systems (CHI) 2019*.

[22] A. Vishwanath, T. Herath, R. Chen, J. Wang, and H. R. Rao, "Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model," *Decis. Support Syst.*, vol. 51, no. 3, pp. 576–586, 2011.

[23] A. A. Cain, M. E. Edwards, and J. D. Still, "An exploratory study of cyber hygiene behaviors and knowledge," *J. Inf. Secur. Appl.*, vol. 42, pp. 36–45, Oct. 2018.

[24] C. Drake, J. Oliver, and E. Koontz, "Anatomy of a Phishing Email."

[25] A. Bergholz, J. De Beer, S. Glahn, M. F. Moens, G. Paaß, and S. Strobel, "New filtering approaches for phishing email," *J. Comput. Secur.*, vol. 18, no. 1, pp. 7–35, 2010.

[26] M. Chandrasekaran, K. Narayanan, and S. Upadhyaya, "Phishing E-mail Detection Based on Structural Properties," *NYS Cyber Secur. Conf.*, pp. 1–7, 2006.

[27] J. Mao, W. Tian, P. Li, T. Wei, and Z. Liang, "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity," *IEEE Access*, vol. 5, pp. 17020–17030, Aug. 2017.

[28] Y. Zhang, S. Egelman, L. Cranor, and J. Hong, "Phinding Phish: Evaluating Anti-Phishing Tools."

[29] S. Sheng *et al.*, "Anti-Phishing Phil: The design and evaluation of a game that teaches people not to fall for phish," *ACM Int. Conf. Proceeding Ser.*, vol. 229, pp. 88–99, 2007.

[30] D. Miyamoto, T. Iimura, G. Blanc, H. Tazaki, and Y. Kadobayashi, "EyeBit: eye-tracking approach for enforcing phishing prevention habits," *Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)* pp. 56–65, 2014.

[31] Y.-Y. Choong, M. Theofanos, K. Renaud, and S. Prior, "Case Study-Exploring Children's Password Knowledge and Practices."

[32] S. Maqsood, S. Maqsood, R. Biddle, and S. Chiasson, "An exploratory study of children's online password behaviours," in *IDC 2018 - Proceedings of the 2018 ACM Conference on Interaction Design and Children*, 2018, pp. 539–544.

[33] D. R. Lamichhane and J. C. Read, "Investigating children's passwords using a game-based survey," in *IDC 2017 - Proceedings of the 2017 ACM Conference on Interaction Design and Children*, 2017, pp. 617–622.

[34] J. C. Read and B. Cassidy, "Designing textual password systems for children," in *ACM International Conference Proceeding Series*, 2012, pp. 200–203.

[35] J. Cole, G. Walsh, and Z. Pease, "Click to Enter: Comparing Graphical and Textual Passwords for Children," vol. 17.

[36] J. Nicholson, L. Coventry, and P. Briggs, "Age-related performance issues for PIN and face-based authentication systems," in *Conference on Human Factors in Computing Systems - Proceedings*, 2013, pp. 323–332.

[37] O. Dele-Ajayi, R. Strachan, A. J. Pickard, and J. J. Sanderson, "Games for Teaching Mathematics in Nigeria: What Happens to Pupils' Engagement and Traditional Classroom Dynamics?," *IEEE Access*, vol. 7, pp. 53248–53261, 2019.

[38] J. C.-Y. Sun, C.-Y. Kuo, H.-T. Hou, and Y.-Y. Lin, "Exploring Learners' Sequential Behavioral Patterns, Flow Experience, and Learning Performance in an Anti-Phishing Educational Game," 2017.

[39] J. C.-Y. Sun and K.-H. Lee, "Which Teaching Strategy is Better for Enhancing Anti-Phishing Learning Motivation and Achievement? The Concept Maps on," 2016.

[40] J. Hardwick, "Top 100 Most Visited Websites by Search Traffic (as of 2019)," *Ahrefs*, 2019. [Online]. Available: https://ahrefs.com/blog/most-visited-websites/. [Accessed: 05-Jun-2020].

[41] Fuse, "The 10 Most Popular Things for Teens in 2019," 24-Jan-2019. [Online]. Available: https://www.prnewswire.com/news-releases/the-10-most-popular-things-for-teens-in-2019-300783768.html. [Accessed: 05-Jun-2020].

[42] K. Tian, H. Hu, D. Yao, and G. Wang, "Needle in a Haystack: Tracking Down Elite Phishing Domains in the Wild" 2018.

[43] J. Cohen, "Statistical Power Analysis," *Curr. Dir. Psychol. Sci.*, vol. 1, no. 3, pp. 98–101, 1992.

[44] Symantec, "Internet Security Threat Report (ISTR) 2019 | Symantec," 2019.

[45] C. D. C. Neil A. Macmillan, *Detection Theory: A User's Guide*. 2004.

[46] C. I. Canfield and B. Fischhoff, "Setting Priorities in Behavioral Interventions: An Application to Reducing Phishing Risk," *Risk Anal.*, vol. 38, no. 4, 2018.

[47] I. K. R. Hatlevik and O. E. Hatlevik, "Students' evaluation of digital information: The role teachers play and factors that influence variability in teacher behaviour," *Comput. Human Behav.*, vol. 83, pp. 56–63, 2018.

[48] S. Michie, M. M. van Stralen, and R. West, "The behaviour change wheel: A new method for characterising and designing behaviour change interventions," *Implement. Sci.*, vol. 6, no. 1, pp. 1–12, Apr. 2011.