How Effective is Anti-Phishing Training for Children?

Elmer Lastdrager and Inés Carvajal Gallardo, ***University of Twente;***
Pieter Hartel, ***University of Twente; Delft University of Technology;***
Marianne Junger, ***University of Twente***

https://www.usenix.org/conference/soups2017/technical-sessions/presentation/lastdrager

This paper is included in the Proceedings of the Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017).

July 12–14, 2017 • Santa Clara, CA, USA

ISBN 978-1-931971-39-3

Open access to the Proceedings of the Thirteenth Symposium on Usable Privacy and Security is sponsored by USENIX.

How Effective is Anti-Phishing Training for Children?

Elmer Lastdrager[1], Inés Carvajal Gallardo[1], Pieter Hartel[1,2], Marianne Junger[1]
[1]University of Twente, The Netherlands
[2]Delft University of Technology, The Netherlands
elmer@lastdrager.com, i.r.carvajalgallardo@utwente.nl, pieter.hartel@utwente.nl, m.junger@utwente.nl

ABSTRACT

User training is a commonly used method for preventing victimization from phishing attacks. In this study, we focus on training children, since they are active online but often overlooked in interventions. We present an experiment in which children at Dutch primary schools received an anti-phishing training. The subjects were subsequently tested for their ability to distinguish phishing from non-phishing. A control group was used to control for external effects. Furthermore, the subjects received a re-test after several weeks to measure how well the children retained the training. The training improved the children's overall score by 14%. The improvement was mostly caused by an increased score on the questions where they had to detect phishing. The score on recognizing legitimate emails was not affected by the training. We found that the improved phishing score returned to pre-training levels after four weeks. Conversely, the score of recognition of legitimate emails increased over time. After four weeks, trained pupils scored significantly better in recognizing legitimate emails than their untrained counterparts. Age had a positive effect on the score (i.e., older children scored higher than younger ones); but sex had no significant influence. In conclusion, educating children to improve their ability to detect phishing works in the short term only. However, children go to school regularly, making it easier to educate them than adults. An increased focus on the cybersecurity of children is essential to improve overall cybersecurity in the future.

1. INTRODUCTION

Fraudsters use phishing to convince victims to give out personal information. Commonly, the fraudsters want credentials that are used to access online services, such as online banking. Even though the impersonated brands that are misused in phishing are predominately financial institutions and payment providers, there has been a recent shift towards retailers and service-oriented companies [3, 4]. Several countermeasures are currently in use to prevent phishing victimization: blocking phishing messages and websites, improving interfaces, and training users [17].

Many training programs have focused on adults (e.g., [27, 5, 1, 18]). An often overlooked group of potential victims is children, with data about children only sparsely available (e.g., in [23]). The current generation of children, sometimes referred to as the *digital generation* or *digital natives*, grew up with the internet. The phrase "digital natives" is being criticized [6], since being a child in this generation does by itself not result in being more digitally capable. Instead, there are lots of opportunities for children, as well as adults, to use technology. Indeed, by the age of nine, many European children have access to the internet [15]. Many of the internet services that adults use, such as social media, email, or online gaming, are used by children as well [7]. A quarter of European children aged 9-10 and 73% of 13 to 14-year-olds have at least one profile on a social media website [15]. In the USA, 68% of teenagers aged 13-14 use social media [24]. Children, and in particular teenagers, are very well represented on the internet, with 92% of American children (13-17 years) [24] and 60% of European children (9-16 years) going online daily [15].

One might wonder why children are at risk. To illustrate why children could be targeted, consider the marketing domain. Marketers know that children have influence over what their parents buy and consequently target children in commercials [10]. In addition to marketing on TV, digital marketing offers even more chances to target children specifically [10, 28]. Phishing is commonly thought to be equivalent to theft of credentials of financial institutions. Since children often don't participate in online banking, what makes them attractive to a phisher? The online footprint of children on social media, websites, and email can be a target by itself. Obtaining access to email or social media accounts is valuable in order to access to a victim's network of friends and family. A phishing message that is sent by a friend is more likely to be opened than one from a stranger [18]. Subsequently, both children and adults within the victim's network can be approached with personalized phishing messages. Alternatively, influencing a child to provide the personal information of his or her parents provides helpful information for a follow-up call or email, even with simple pieces of information such as a phone number or home address. Training is needed to reduce the risk of initial victimization. Just like adults, children need to develop the ability to identify fraudulent communication, such as phishing emails.

Anti-phishing training can be administered in various ways. Advice can be given on an individual level, such as parents

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

teaching their child how to ride a bike. Alternatively, one may educate a group at the same time; for example, schools teach skills like arithmetic to entire classes. When possible, educating a group of children can be more efficient. Since most children attend school, they are used to getting information in a class setting. Furthermore, when parents are insufficiently experienced to educate their children in the area of cybersecurity, this topic should be taught at school.

Education tackles only a part of the problem. An important issue is knowledge retention. One of the difficulties with user training is the extent to which the audience remembers the lessons over the long term. Retention indicates the effectiveness of training. Additionally, it is important to know how often to repeat training. This is true for traditional training, as well as alternative methods of creating user awareness, such as training by playing games [22, 34]. Studies performed on adults found no significant decay in performance from one week up to one month after the intervention [23, 21, 1, 27, 20]. This suggests that improvement of awareness after training is retained in the relatively short term. The question arises whether the same applies to children, as well as, more importantly, whether the improvement in awareness is stable over a longer period.

Children are very active online and can be the target of phishing e-mails. Accordingly, like adults, they should be trained to reduce the risk of victimization. This raises three questions to be answered. Firstly, what are children's abilities to detect phishing emails and websites? Secondly, what effect does cybersecurity training have on the children's ability to detect phishing? Thirdly, after receiving an awareness training, how well do children retain this knowledge? To answer these questions, we conducted empirical research.

Our contributions are: (1) to our knowledge, we are the first study to focus on the effect of anti-phishing training on children; (2) the training was based on storytelling and resulted in an improved detection of phishing in the short term and an improved detection of legitimate messages after 2–4 weeks; (3) we show that subjects with more online exposure, as well as older children, score better on a phishing identification test.

2. METHODOLOGY

An experiment was conducted at six schools in the Netherlands, using a cybersecurity training program that was designed for children aged 9–12. We tested their ability to recognize phishing and measured the effect of an intervention.

2.1 Design & Concepts

The experiment used a 2x2 between-group design. The training intervention was given on a group level (i.e., in a classroom), and we wanted to preserve the anonymity of the pupils. Therefore, no identifying information was recorded on the tests. Consequently, we did not record demographic data other than age and sex. The independent variables were the experimental condition (intervention or control) and the retest duration (measured in number of weeks). The outcome variable is the score on the test, ranging from 0 (no correct answers) to 10 (all answers correct). Five other variables were recorded to identify differences between groups and measure for certain individual differences: sex (male/female); age; possession of email address (yes/no); possession of a Facebook account (yes/no); and whether the subject had received a phishing email before (yes/no/unknown).

We will briefly discuss why these variables were included. Firstly, the subject's sex (male/female) was recorded because several phishing studies found that men are less prone to phishing victimization than women [5, 33, 18, 23], though other studies found no relationship [2, 13, 25]. Age was recorded with the expectation that older subjects would outperform younger ones [23, 33, 2]. Finally, the Routine Activity Approach states that for a crime to occur, a target and an offender must converge in the absence of a capable guardian [12]. Consequently, we expected children who are more active online to be more exposed to phishing. Therefore, subjects were asked whether they possess their own email address and Facebook account, and whether they have received a phishing email in the past.

In this paper, we use the terms "children" or "pupils" interchangeably to refer to the subjects of the study. "Teacher" refers to the school teacher of the pupils. The trainer is a researcher performing the study (by giving the presentation).

To establish the effectiveness of the cybersecurity training, we formed two types of groups: *intervention* and *control*. The intervention group was made up of school classes that received the cybersecurity training, followed by a capability test. To evaluate the effectiveness of the training, we compared the intervention group with a control group that received training after the study was finished (see Section 2.2).

2.1.1 Training and Procedure

A cybersecurity training program was developed for this experiment, consisting of an interactive presentation and a test. During the 40-minute presentation, the trainer would introduce and discuss cybersecurity with a class of pupils. The trainers were researchers and master's students specializing in cybersecurity. Asking children for their attention during a presentation can be challenging. Storytelling is an efficient method for non-experts to share in an expert's knowledge [31]. Therefore, the trainer used short stories and examples focussed on children to attract their attention.

The presentation provided the children with the necessary means of recognizing cyber misbehavior and advice on what to do. Topics included cyberbullying, hacking, phishing and identity theft. For phishing, we first explained what phishing is. Then, we showed an educational TV commercial that had been designed by the Dutch banking association [37]. Following the commercial, we asked the children in a group discussion what clues one should look for. Afterwards, we introduced four clues for identifying phishing emails: (1) how to find a URL from a hyperlink and how to assess where a URL leads to; (2) grammar, spelling, and the general type of language used; (3) presence of a sense of urgency or use of threats; and (4) the sender address. Furthermore, we showed two clues for websites: (1) the URL and (2) the need for an HTTPS connection when entering any data. During the training, the children were given ample opportunity to tell about their experiences, which helps the attendees remember the message. This led the children to share their own advice on how to prevent victimization, along with the advice that was included in the training. The trainer informed the

children about the effectiveness of their own advice. Where needed, alternative advice was provided.

During the experiment, researchers went to schools in pairs. There were several practical constraints in time and availability. For example, schools had to book time to receive us, so there was a strict requirement to finish in time. Within classes of the intervention group, the trainers gave a presentation to the pupils. After the presentation, the children were given a paper-based phishing awareness test. Classes in the control group were only given the phishing test. No further explanation was provided, other than that the trainers would be back at a later time. Some pupils asked questions about a particular part of the test. The trainers answered that the pupil should pick the answer that made the most sense to the pupil.

After several weeks, each class was visited again. All pupils were given another paper-based phishing test. Finally, each child was given a one-page debriefing letter that explained and summarized the study. Additionally, all subjects were encouraged to discuss the test with their parents and contact one of the researchers with any questions.

2.1.2 Testing

Establishing the ability of children to detect phishing was measured using a paper-based phishing test. The participating schools did not have a computer available for each pupil. To allow school participation with the least effort, we chose a paper-based test over a computer-based test. The method of testing phishing ability and the introduction to the test can influence the results. For example, Parsons et al. [29] have shown that primed study participants are significantly better at discriminating between phishing and non-phishing compared to uninformed participants. To reduce this bias, children were not told that the goal was to discriminate phishing from non-phishing. Rather, the test was introduced as a 'cybersecurity test.'

The phishing test consisted of 10 questions, with six emails and four websites to judge. Both legitimate and phishing emails and websites were included. One correct answer was worth a point, and number of correct answers was the student's score on the test. Answering everything wrong would give a score of 0; answering everything correctly gave a 10. For each email or website in the test, a decision had to be made whether or not to take action. Although it was not stated explicitly, the pupils made a phishing or not phishing decision. Participating pupils were asked to note what kind of action they would take. Subjects' scores can vary depending on the type and origin of emails they have to judge [29]. Therefore, diversity in the types of emails and websites is essential to obtain a valuable result. Each question contained a clue as to why it should or should not be trusted. Some clues were explicit, such as a wrong link in an email or an unusual sender address. Others were based on the content, such as expressing urgency and spelling errors. For content-based clues, we made sure to include several in an email or website. All clues were mentioned in the training. The questions, emails, and websites were tailored to children and included a variety of different companies, such as toy stores, TV programs, game websites, a bank, and social media. The questions were not based on real-life phishing emails, since we are unaware of phishing attacks that target children specifically. However, we used existing legitimate emails and websites and adapted them, just like a phishing offender would do.

The tests were aimed at measuring the ability to identify emails and websites as phishing or legitimate correctly. However, using the same phishing test for the initial measurement as well as the re-test could result in the subjects remembering the questions. To avoid this memory effect, three sets of questions were used to measure the ability of children to detect phishing emails and websites. Three versions of the test were made: A, B, and C. Tests A and B included a front page with questions about the online exposure of the subjects. Test C was used in the pilot phase of the experiment and contains reordered questions from Test A.

Each subject got an overall score, the outcome variable. However, human beings generally assume that a message is truthful, and have great difficulty recognizing lies [26]. This has been called the truth bias [19, 26, 9]. We need to consider two parts in the subjects' performance: detecting lies (phishing) and detecting truth (legitimate). To do so, we made two equal-sized sets of questions. One set contained phishing, the other contained legitimate communications. By separately grading both sets of questions, we could distinguish between the ability to detect lies versus the ability to detect the truth. The overall score of a subject was calculated as the sum of both sets. For example, if a subject scored 3.0 out of 5 for recognizing phishing, and 2.5 out of 5 for recognizing a legitimate communication, the overall score would be 5.5 out of 10.

2.1.3 Retention

To measure knowledge retention, each school class took two phishing tests to test their ability to recognize phishing over time. Classes in the intervention condition received the training, followed by a test. Immediately after groups in the intervention condition finished their tests, the correct answers were discussed in class. This allowed the children to ask questions once more and get feedback on their decisions, thereby increasing the learning effect. After either 2 weeks (14 days), 4 weeks (28 days), or 16 weeks (64 days) a second test was done. Classes in the control condition did one test initially, followed by a re-test after 2 or 4 weeks. For the control condition, the results of the tests were not discussed in class. Unfortunately, classes in the control group that were scheduled for a re-test after 16 weeks were unable to participate the second time. This makes it impossible to compare the intervention group with a control group at 16 weeks. Therefore, our analysis will focus on the retention between 0 and 4 weeks.

2.2 Ethics

As with any experiment with humans, ethics are important. First of all, the design of this study was approved by the institutional review board of the University of Twente. The study was designed such that the subjects were not hurt or distressed in any way. Furthermore, each participating school was asked for permission to conduct the training and test their pupils. Additionally, we asked each participating school to distribute informed consent letters to the parents of their pupils. Parents were asked to sign and return the informed consent, either to the school or by email to the researchers. The contact information of the researchers was

included in the informed consent, in case parents had questions. Several parents contacted the researchers. Only when the parents of a pupil had signed the informed consent and returned this to the school could a child participate as a subject.

After finishing the experiment, each subject was given a debriefing letter. The letter was written for the child and encouraged him or her to discuss the training with his or her parents. Furthermore, the contact details of the researchers were included in the debriefing, in case anyone had questions. After finishing the experiment, nobody contacted the researchers with questions.

From the point of view of the experiment, it was important to separate intervention and control groups. We considered it unethical to deprive subjects in the control group of a cybersecurity training. Therefore, after finishing their second phishing test and concluding their participation as subjects, pupils in the control group received the training too.

2.3 Setting

The experiment was held at six schools in the Netherlands, of which five primary schools and one secondary school. Each participating school gave permission for two sessions for at least one class. Every class received two tests (of 20-30 minutes each), and one intervention (about 40 minutes). Classes were randomly assigned to either an intervention group or a control group, and were additionally assigned a retention period by the researchers. All tests were taken individually by the subjects. The researchers were present to answer questions, but would never give away the correct answer. The subjects were told to answer what they would do if they had received the email or visited the website.

2.4 Subjects

The subjects were 353 pupils from six participating schools. All subjects were aged between 8 and 13 (M=10.66; SD=1.05), and over half (54%) were female. Children could join the training only if their parents had given their written consent before the start of the program (refer to Section 2.2 for more information). Children who did not have permission from their parents were temporarily sent to another classroom. If changing rooms was not possible, non-participating children were moved to another part of the same classroom to work on another task. Each child was assigned to an intervention or control group, based on the class they were in. This resulted in 181 children in the intervention group who received training, compared with the control group consisting of 172 children. The re-test was taken by 177 children. We included the week 0 data for several classes that were unable to participate for the re-test. Specifically, the missing classes consisted of all control group classes for the 16-week re-test. This resulted in the exclusion of the 16-week intervention group's re-test, since we could not compare them with their control group counterparts. Therefore, the number of subjects in week 0 is significantly higher compared to those for the re-tests in weeks 2 and 4. The exact number of subjects at each stage in the experiment is listed in Table 1.

Table 1: Number of subjects in each stage of the experiment.

Group	Week 0	Week 2	Week 4
Intervention	181	49	38
Control	172	32	58

2.5 Analysis

The three research questions guided the analysis. Descriptives of the control groups provided an answer to the first research question (i.e., what are the children's abilities to detect phishing emails and websites?). Furthermore, we tested whether the subject's characteristics influenced the score. An independent group t-test was used to measure the effect of the subject's sex and possession of an email account. The second research question was: what effect does cybersecurity training have on children's ability to detect phishing? To measure this effect, we compared the intervention group and the control group at 0 weeks. This was done using an independent group t-test, showing the difference between trained children (the intervention group) and untrained children (the control group). The third research question quantified the retention of the training. To answer this question, several linear regression models were developed. Firstly, a multi-level model was tested, measuring whether the school attended by the subject accounted for the results of the pupils. Even though the multi-level model was significant, the intraclass correlation was low (i.e., below 0.025). Therefore, linear regression was used instead. We developed several such models.

Model I uses the type of experiment (i.e., intervention or control), the number of weeks, and the interaction of these two as the predictors. ExperimentType shows the effect of the training on the score. The number of weeks indicates retention over time. Additionally, it is interesting to learn whether the effect of the training increases or decreases over time. For example, teaching someone a skill such as biking results in a higher level of skill over time if the person practices on his or her own. Therefore, the interaction between having participated in the intervention and the number of weeks (ExperimentType $\times$ Weeks) was taken into account as well. With this interaction, we could analyze whether the intervention resulted in better results as time progressed. A second model including social variables was constructed as Model II. Age and sex were added to the variables from Model I. Age was included since related literature suggested that older subjects score better than younger ones. The literature is inconclusive when it comes to sex and phishing victimization. Therefore, we added sex as a variable. Finally, Model III combines Models I and II and adds the test version and school, to show their potential influence on the overall score of the subjects. The school and test version variables were moderately correlated (r=0.68), as a consequence of Test C being used only in the pilot of the study. This results in collinearity in the model. Therefore, we omitted Test C from the model. These three models were used to predict the subject's overall scores on the tests.

Using the overall score as a measure of the ability to recognize phishing from legitimate is by itself insufficient. As discussed before, one needs to distinguish the differences in the scores of recognizing phishing and recognizing legitimate communications. To accommodate this, additional models were developed to distinguish lie detection and truth detection in the analysis. This lead to the introduction of six

models. Phish-I through Phish-III were based on the previously described models I-III, but used the phishing (lies) score instead of the overall one. Additionally, Legit-I to Legit-III were developed to model the scores of the legitimate (truth) questions.

3. RESULTS

The first research question concerned the ability of children to detect phishing. This translates to the scores of the control group at the beginning of the experiment, at week 0. The average overall score of this control group is a 6.02 (Table 2) on a scale from 0 to 10. The overall score consisted of two parts: phishing (0–5 points) and legit (0–5 points). When considering only the questions that were related to phishing, the control group scores 3.74 on average, with a 95% confidence interval of [3.62, 3.88]. The mean score for labeling legitimate questions as such was lower: 2.26 (95% CI [2.09, 2.44]). In addition to the average scores of the control group, we also measured the effects of several subject characteristics on the overall score for all subjects. There was no significant effect of sex on the score, indicating a lack of evidence that boys performed differently from girls ($t(633) = -0.62$, $p=0.53$). There was a significant effect of age on the score, with older pupils scoring higher than younger ones ($F(1,633) = 6.28$, $p=0.01$, $R^2=0.010$, Adj. $R^2=0.009$). The effect of the school on the subject's score was significant ($F(5,636)=7.54$, $p<0.001$, $R^2 = 0.056$). One school scored significantly lower compared to the others ($B=-0.80$; $p=0.004$). Most of the subjects (80.3%) indicated having their own email address. Having one's own email address significantly influenced the score, with subjects having their own email address performing better than those without ($t(469)=3.68$, $p<0.001$). On the topic of social media, 26.6% of the subjects indicated having their own Facebook profile. Subjects with their own Facebook profile scored significantly higher than those without a Facebook profile ($z=2.330$, $p=0.02$, $r=0.10$). Thirdly, when asked whether they had ever received a phishing message, 8.9% answered 'yes', 37.4% answered 'no' and the remaining 53.7% responded that they did not know. Whether or not the subjects received a phishing email before was not significantly related to the subject's score ($F(2, 468) = 0.61$, $p=0.55$). A subject's online exposure did result in higher odds of having received a phishing message before ($F(2,215) = 6.25$, $p=0.002$, $R^2=0.040$), whereby having an email address was a significant indicator ($B=0.16$, $SE=0.05$, $p=0.04$).

To answer the second research question, the effect of the training was measured. Since three paper-based phishing tests were used in the experiment, we wanted the results to be comparable regardless of the version of the test. The mean overall results of pupils taking different tests were not significantly different from each other: A and B ($t(470)=1.89$; $p=0.059$); A and C ($t(307)=0.98$; $p=0.326$); B and C ($t(451)=-1.214$; $p=0.225$). Figure 1 shows the differences in scores in three box plots. The means and confidence intervals under all experimental conditions are listed in Table 2. The training itself resulted in an improvement in the scores of the participants in the intervention group that was statistically significant compared to the control group ($t(634)=-10.56$, $p<.001$). The effect size was $r=.39$, indicating a medium-sized effect [11]. In comparison, if we include only the first measurement (i.e., week 0), there is a significant difference between the untrained and the trained children as well ($t(351)=-5.19$; $p<0.001$). The training in week 0 had a small effect size of $r=.27$. These results show the effectiveness of adding a simple and short cybersecurity training to the curriculum of schools.

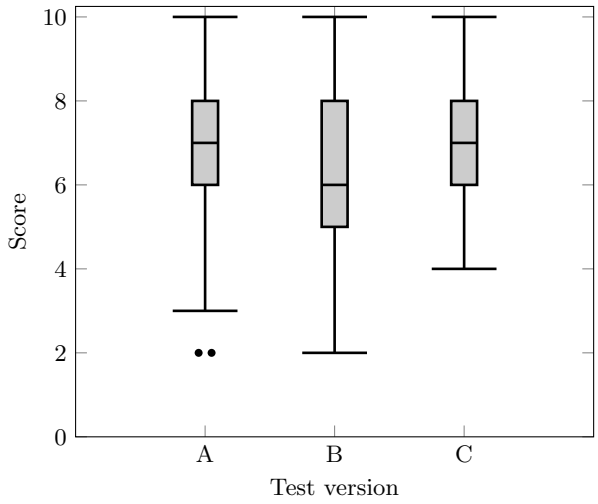


Figure 1: Box plot of three phishing tests of all observations (N=636).

To answer the third research question, retention over time was measured. Several linear regression models were constructed, the results of which are included in Table 3. Model I shows the influence of the cybersecurity training intervention on the score, as well as the effect over time, while controlling for the interaction effect. The resulting Model I is significant and explains 18.6% of the variance ($F(3,526) = 41.77$, $p<0.001$). Model II adds social predictors to Model I, resulting in a model that explains 19.8% of the variance ($F(5,523) = 27.63$, $p<0.001$). Finally, Model III includes the school as well as the version of the test, as well as the predictors from the other models. Model III is significant and explains 25.7% of the variance ($F(11,517) = 17.46$, $p<0.001$). In all three models, the effect of training significantly influenced the score of the subjects throughout the following weeks ($\beta=0.23$, $p<0.001$). Furthermore, the intervention group score significantly higher over time compared to the control group. Figure 2 plots Model III based on the number of weeks passed, split into intervention or control group, to show these effects visually.

To measure the differences in detecting lies from detecting truth, we developed additional models based on Models I, II and III. Instead of using the overall score as the outcome variable, we used the phishing score or the legitimate score, respectively. Since half of the questions were phishing, the scores range from 0 (all wrong answers) to 5 (all correct). Models Phish-I to Phish-III use the score of recognizing phishing. The model results can be found in Table 4. Model Phish-I includes the same predictors as the normal Model I, and is significant and explains 8.3% of the variance ($F(3,526)=15.36$, $p<0.001$). Model Phish-II is significant and explains 8.3% of the variance as well ($F(3,523)=9.26$, $p<0.001$). Model Phish-III is significant as well and explains 13.1% of the variance ($F(11,517)=9.60$, $p<0.001$). Compared to the models of the overall scores, different effects emerge. For example, subject age and weeks since inter-

Table 2: Mean score and 95% confidence interval per experimental setting.

		Overall Score		Phishing Score		Legitimate Score	
Type	Week	Mean	95% CI	Mean	95% CI	Mean	95% CI
Cont	0	6.02	5.79–6.26	3.61	3.45–3.77	2.41	2.20–2.62
Exp	0	6.87	6.65–7.09	4.26	4.15–4.38	2.61	2.41–2.80
Cont	2	5.72	5.21–6.23	4.09	3.74–4.45	1.62	1.17–2.08
Exp	2	7.95	7.58–8.34	4.33	4.12–4.53	3.63	3.28–3.99
Cont	4	6.14	5.75–6.53	3.96	3.70–4.23	2.17	1.79–2.55
Exp	4	8.13	7.67–8.60	4.00	3.73–4.27	4.13	3.81–4.46
Cont	all	6.01	5.82–6.20	3.74	3.62–3.88	2.26	2.09–2.44
Exp	all	7.35	7.19–7.51	4.23	4.15–4.32	3.11	2.97–3.26

Table 3: The linear regression models of the overall score.

	Model I			Model II			Model III		
Characteristic (reference)	B	SE B	β	B	SE B	β	B	SE B	β
ExperimentType (control)	0.92***	0.16	0.28	0.90***	0.16	0.27	1.00***	0.12	0.10
Weeks	0.01	0.06	0.01	0.03	0.06	0.03	0.11	0.12	0.10
Weeks × ExperimentType	0.34***	0.08	0.23	0.36***	0.08	0.24	0.30**	0.08	0.20
Age				0.18**	0.06	0.11	0.19**	0.07	0.12
Sex (female)				0.08	0.13	0.02	0.14	0.14	0.04
Test version (A)†									
– Test B							-0.17	0.39	-0.05
School (1)									
– 2							0.89**	0.33	0.16
– 3							0.44	0.31	0.08
– 4							-0.33	0.34	-0.05
– 5							0.30	0.43	0.07
– 6							-0.24	0.47	-0.07
Constant	5.99***	0.11		4.04***	0.69		3.80***	0.85	
R^2	0.186			0.198			0.257		
Model significance	0.000***			0.000***			0.000***		
N	530			529			529		

Note. Coefficients unstandardized (B) and standardized (β). SE=Standard Error. Significance (χ^2): * p<0.05; ** p<0.01; *** p<0.001. †Due to collinearity, the output of Test C was omitted.

vention in Phish-III are not significant, whereas they are in the overall Model III. The differences are more easily viewed when Model Phish-III is plotted in Figure 3a. At week 0, the intervention group's scores differ significantly from the control group, as shown by the confidence intervals. However, in week 4, there is no significant difference between the intervention group and the control group anymore. The control group scored similarly in week 4 compared to week 0. Subjects within the intervention group scored significantly lower in week 4 compared to week 0.

In additional to the three phishing-only models, three legit-only models were constructed. Similarly, three models, Legit-I to Legit-III were constructed based on the overall Models I to III, respectively. The results of these models can be found in Table 5. Model Legit-I was significant and explained 15.1% of the variance (F(3,526)=42.57, p<0.001). Model Legit-II was significant and explained 16.4% of the variance (F(5,523)=29.59, p<0.001). Model Legit-III was significant and explained 26.0% of the variance (F(11,517) =20.28, p<0.001). A graph showing Model Legit-III is included in Figure 3b, with scores ranging from 0 to 5 for all five questions testing legitimacy. There are no significant differences in score at week 0 between the intervention group and the control group for the legitimate scenarios (z=-1.17; p=0.24). In week 4, however, the scores of the intervention group and control group differ significantly (z=-5.85; p<0.001). During the experiment, the score of the control group did not change significantly (t(228) = 1.11; p=0.27). In the intervention group, a significant increase in score was observed between week 0 and week 4 (z=-6.05; p<0.001).

4. DISCUSSION

The concept of testing the ability to detect phishing in an educational setting is challenging [32]. Getting the attention of children aged 8–13 to focus on cybersecurity is no less of a challenge. Untrained children are mediocre at discriminating phishing emails and websites from legitimate ones, scoring 6.02 out of 10 in our experiment. However, subjects trained in a single 40-minute training session and interactive discussion scored 6.87 out of 10, an increase of 14% over their untrained peers. The overall score by itself is not sufficient as a measurement of accuracy, since humans are generally not very good at recognizing lies [26]. Therefore, we distinguished the correctness scores for phishing and legitimate questions.

Table 4: The linear regression models of the phishing-only score. The construction of the models is similar to Table 3.

	Model Phish-I			Model Phish-II			Model Phish-III		
Characteristic (reference)	B	SE B	β	B	SE B	β	B	SE B	β
ExperimentType (control)	0.65***	0.10	0.34	0.65***	0.10	0.34	0.70***	0.10	0.37
Weeks	0.10**	0.04	0.16	0.10**	0.04	0.16	0.02	0.07	0.04
Weeks × ExperimentType	-0.15**	0.05	-0.18	-0.15**	0.05	-0.17	-0.18**	0.05	-0.22
Age				0.01**	0.04	0.01	0.05	0.04	0.05
Sex (female)				-0.00	0.08	-0.00	0.09	0.08	0.05
Test version (A)†									
– Test B							-0.21	0.24	-0.10
School (1)									
– 2							0.56**	0.21	0.17
– 3							0.08	0.21	0.03
– 4							0.20	0.22	0.06
– 5							0.95**	0.27	0.41
– 6							0.77**	0.29	0.40
Constant	3.63***	0.08		3.50***	0.44		2.59***	0.52	
R^2	0.083			0.083			0.131		
Model significance	0.000***			0.000***			0.000***		
N	530			529			529		

Note. Coefficients unstandardized (B) and standardized (β). SE=Standard Error. Significance (χ^2): * p<0.05; ** p<0.01; *** p<0.001. †Due to collinearity, the output of Test C was omitted.

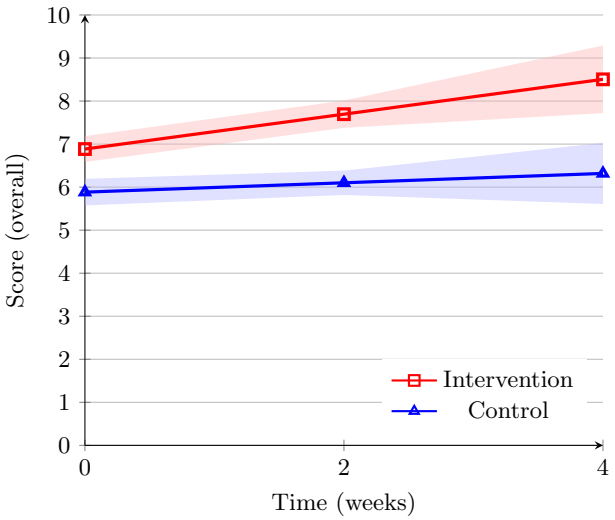


Figure 2: Overall predicted ability scores over time, in number of correct answers (0–10). Shades indicate 95% confidence interval. N=529.

We found that training improved the ability to recognize phishing directly following the training, but it did not significantly change the ability to identify legitimate emails correctly. This phenomenon has been discussed in the literature. Hauch et al. [16] have shown in a meta-analysis that training improves both overall accuracy and lie detection, but not truth detection accuracy. This was also the case in our experiment; the subjects did not score significantly better on truth accuracy of legitimate emails and websites on the test directly following the training, compared to the control group. This can be explained by the focus of our training on how to detect phishing. According to Hauch et al. [16], if the focus of training is on deception detection, the subject's post-training truth accuracy remains unaffected. An alternative explanation would be that the training made the subjects paranoid. However, if that were to be the case, the subjects would have to score lower on recognizing legitimate emails, which was not the case.

The overall scores of trained subjects improved significantly over time, indicating a good knowledge retention of the subjects. Within the control group, the overall scores remained stable. When considering only the phishing questions, subjects from the intervention group suffered from a small decay in their ability to recognize phishing. Specifically, after 4 weeks, the ability of the intervention group to recognizing phishing matched the level of the control group. Regardless of the decay over time, the scores on the phishing questions were relatively high, with averages of correct answers between 3.7 and 4.4 questions. Since 5 was the maximum, we believe that there is a ceiling effect: many subjects achieved the highest score, and could not improve their scores further. Our test consisted of 10 questions composed of two sub-tests, five legitimate and five phishing. This means that subjects could not receive higher scores than 5 on both subtests, which is the maximum on our measures. When many subjects have the maximum score, their scores cannot be distinguished. Figure 3b illustrates this clearly for the intervention group. Therefore, only less-performing subjects could increase their score after training. The subsequent score decay over time shows that the effect of the training, in terms of the ability to recognize phishing emails, fades within a month. To the best of our knowledge, no similar phishing tests have been undertaken with children, making comparisons with other phishing literature difficult. There are studies on phishing interventions with adult subjects, which found no significant decay of the trained subject's abilities after 7 to 28 days [23, 21, 1, 27, 20]. However, there are major methodological differences, since the abovementioned studies use interactive, computer-based methods of training, such as playing games [23, 21, 27] or roleplay-

Table 5: The linear regression models of the legitimate-only score. The construction of the models is similar to Table 3.

	Model Legit-I			Model Legit-II			Model Legit-III		
Characteristic (reference)	B	SE B	β	B	SE B	β	B	SE B	β
ExperimentType (control)	0.27	0.14	0.09	0.25	0.14	0.09	0.30*	0.14	0.10
Weeks	-0.08	0.05	-0.09	-0.07	0.05	-0.07	0.08	0.11	0.09
Weeks × ExperimentType	0.49***	0.07	0.38	0.51***	0.07	0.39	0.48***	0.07	0.37
Age				0.17**	0.06	0.11	0.14*	0.06	0.10
Sex (female)				0.08	0.12	0.03	0.05	0.12	0.02
Test version (A)†									
– Test B							0.04	0.35	0.01
School (1)									
– 2							0.33	0.29	0.07
– 3							0.36	0.26	0.07
– 4							-0.54	0.29	-0.10
– 5							-0.65	0.40	-0.18
– 6							-1.02*	0.41	-0.35
Constant	2.36***	0.11		0.54	0.62		1.21	0.74	
R^2	0.151			0.164			0.260		
Model significance	0.000***			0.000***			0.000***		
N	530			529			529		

Note. Coefficients unstandardized (B) and standardized (β). SE=Standard Error. Significance (χ^2): * p<0.05; ** p<0.01; *** p<0.001. †Due to collinearity, the output of Test C was omitted.

ing [1]. However, within the field of social engineering, it has been reported that an intervention to increase awareness is subject to significant decay [8], showing social engineering awareness returning to pre-intervention levels after two weeks.

While the phishing score decreased over time, the score for legitimate questions followed a rather different pattern. The score over time increased significantly, contrary to our expectations. After two and after four weeks, subjects in the intervention group were able to correctly recognize legitimate scenarios significantly better than subjects from the control group. The cybersecurity training may have triggered the interest of the children, causing them to pay more attention to messages they receive, or to think about the lessons learned. Another possible explanation is that the subjects trained themselves based on emails they received in their daily lives. This may be compared to learning how to ride a bike, where an initial set of skills and knowledge is needed to start biking, and with more practicing, performance increases over time. In other words, training made the children look more closely at the emails they received, after which they were better at identifying legitimate emails.

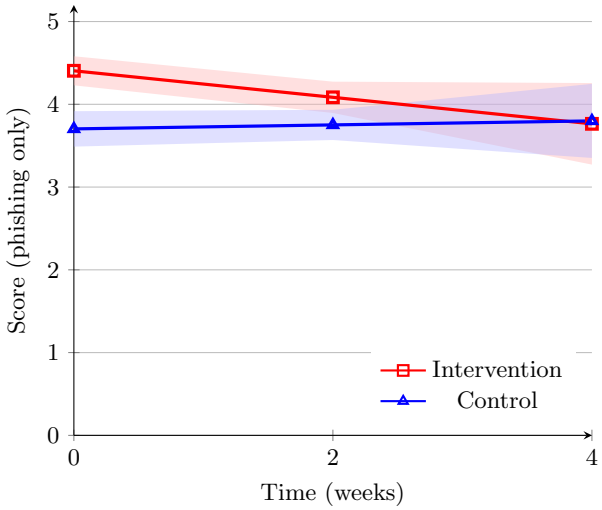
Further trainings, sometimes called boosters, could be used to increase these abilities and counter decay of the ability to recognize phishing [20, 30]. However, regular training is costly. In the context of children, it may be infeasible for schools to introduce boosters on a regular basis. This is especially the case when the retention of knowledge is short (i.e., a month). Training using different methods, such as letting the subjects play a game [23, 21], may be less affected by this disadvantage since the subjects can play the game regularly without supervision. Before introducing additional training, however, better measurements should be used to identify the problem better. One possible fix is an extensive test with more questions and more challenging questions, which could be used to avoid a possible ceiling effect. That way, subjects would be less likely to get the maximum score, and decay or increase effects should be more visible.

Another finding is that older children score better than younger ones. This is in line with similar studies about phishing interventions on adults. In several studies, young adults perform worse than older ones [33, 2]. In particular, a large-scale study [23, 33] found that teenagers between 13 and 17 perform worse than adults in phishing tests. A possible reason for this result is lower education and fewer years of internet experience [33]. Furthermore, subjects in this study who have their own email address or a Facebook profile scored significantly higher than other subjects. This suggests that, indeed, internet experience may be an influential factor. Another factor that could influence the subject's score is the training itself. Despite efforts to make all trainings similar, there are group dynamics involved, especially when relying on interaction (e.g., stories) with the subjects.

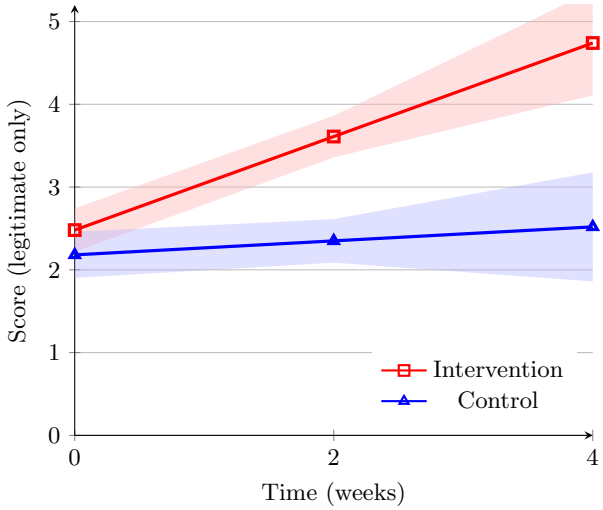
Other candidate relations did not significantly contribute to the final score of a child. In particular, the sex of the child had no significant influence on the score, when controlling for other variables. Specifically for children, sex differences are not necessarily to be expected at all. For example, boys only begin to take more risks than girls between the ages of 9 and 11 [35]. The lack of differences could be explained by the age groups of the children that participated. Additionally, even for adults and adolescents, the existence of a relation between sex and phishing knowledge is doubtful in existing literature [2, 13, 25]. The interaction between age and sex did not predict phishing knowledge of children either.

4.1 Limitations

There are several limitations to the results of this study. Even though the intervention condition was given per class, this did not prevent children in one class from talking to their peers in other classes. Since all parents were informed and asked for permission beforehand, they could have dis-



(a) Includes only the phishing questions.



(b) Includes only the legitimate questions.

Figure 3: Predicted ability score split by phishing and legitimate. Shades indicate 95% confidence interval. N=529.

cussed the topic of cybersecurity with their children. Unknown external factors may be responsible for the increase over time. For example, the participating children may have seen one of the phishing awareness commercials on TV. Personal experience of the researchers was that indeed one of of these three explanations was plausible. One of the colleagues at the University of Twente, who was not involved in the study, had a child in the intervention condition. The colleague mentioned that his children and the other parents were enthusiastic about the intervention and that he had talked about it at home. This example could explain the increase in ability over time that was observed. Moreover, this colleague had other children in the same school. Hence, the intervention could have influenced children in the control condition. However, we do not see indications of that effect in the data.

A possible critique on the study is that the children know that they are being tested. The results, therefore, do not necessarily reflect their ability when receiving an email in the wild. While this is true, we consider the tests an appropriate way to measure the subject's ability to recognize phishing. The subjects' scores are arguably different from how they would respond to a phishing email in their own inbox, since more factors are involved. Factors such as language (an eight-year-old Dutch child receiving an English email) and expectancy (not having a bank account) could increase their real world score. On the other hand, factors like attention (doing other things in parallel) and limited interfaces (not being able to check the link on a tablet computer) could affect resilience in the real world. Furthermore, the subjects received a second test a period of time after the first. This means that they know what to expect when they start the second test.

This study may suffer from an assignment bias. Even though the groups were assigned at random to one of the conditions, the number of schools that participated is limited. Furthermore, all schools are located in two cities in the east of the Netherlands. The results might be affected by factors unknown to the researchers. A nation-wide study on randomly selected schools could counter such biases regarding region and quality of teaching.

A presentation (or lecture) is one way to deliver a message to pupils. Other ways of teaching may be more efficient, such as using games [14]. We chose a traditional presentation-based intervention because it is relatively simple to apply to current primary schools. The pupils do not need to have access to a computer, and a presentation and paper-based test fit in well with the rest of the daily program and activities. Alternatively, game-based anti-phishing solutions [22, 34] may yield better results and could have different retention properties.

Using a paper-based test with images raises questions regarding the representativeness of the resulting score compared to real-world phishing. Whereas using static images or screenshots is not optimal, they have been used before in phishing experiments [36, 29, 33]. We believe there is little difference between seeing an image on a screen or seeing one printed on paper. Furthermore, not all subjects may be equally computer literate, and using static images on paper results in a level playing field.

Finally, all students filled in the tests anonymously. Therefore, no repeat measurements were available at an individual level. The analyses could therefore not be performed on repeated-measures samples. Rather, we treated the test results as independent samples. As a consequence, the reported results are conservative and an underestimation, as they miss the power of a repeated-measures test.

5. CONCLUSIONS

Children need to understand digital risks to reduce the risk of victimization on the internet. Understanding digital risks is important for children as well as adults. However, the majority of children are self-taught when it comes to the internet [7], making it unlikely they will systematically learn how to act safely. To learn about the abilities of children in detecting phishing emails and websites, researchers had

children aged 8–13 take in a phishing recognition test. Half of the children received training before the test, and the other half did not. Both trained and untrained children were tested for the ability to distinguish phishing emails and websites from legitimate ones. Several schools participated in the study. A first indicator of the practical need for such training arose while performing the experiment. During the training, as more pupils started sharing their stories, they became very enthusiastic and asked lots of questions. In most classes, at least one child knew a phishing victim. These victims were mostly relatives or neighbors. The most common situation in the stories that were told was a victim losing money due to filling in banking credentials on a phishing website. Hearing stories from their peers impacted the children and provided them with a warning message stronger than the presenters could ever give.

Until novel anti-phishing techniques are developed and deployed on a large scale, user training seems to be important. For adults as well as children, that means creating an improved knowledge of the subject for as many individuals as possible. In many countries, all children aged 9 or older attend some form of education. Potentially, this makes it feasible to embed a cybersecurity training in their curriculum, effectively training the entire population of children.

In our experience, both schools and parents are very willing to embed lessons about cybersecurity in the curriculum. Our request to give a training was well received. In particular, incidents with phishing, cyberbullying, and other cyberthreats are often in the news. Teachers and parents reported being worried about those issues. At the same time, teachers at schools where we gave a training, found the course highly informative for themselves as well. Techniques for establishing the validity of an email were unknown to them. Several teachers mentioned that hovering over a hyperlink or checking the sender address were valuable approaches for them. Training teachers should, therefore, be the first step in cybersecurity education. Where needed, universities and practitioners (e.g., IT security firms) could provide help. There are existing initiatives, such as the $(ISC)^2$ Safe and Secure Online[1] where security professionals visit schools. Such initiatives should be extended to more countries and expanded in size, and new ones should be developed.

Training children increased their short-term ability to distinguish phishing from legitimate correctly. Specifically, their ability to recognize phishing increases significantly after an in-class training. However, this increased ability is subject to decay. After four weeks, the ability to recognize phishing for trained children diminished to the level of their non-trained counterparts. This suggests that the training created knowledge, but that this knowledge only lasted through the short term. On the positive side, trained children did continue to perform better in recognizing legitimate emails as such. This increases the odds of legitimate communications reaching the end user. Increasing the ability to recognize phishing requires good awareness.

All in all, we believe that researchers and practitioners in the field of cybersecurity should not only focus on adults, but that material for children should be developed in parallel. Phishing, specifically, is too often seen as an adult-only crime. The children of today are the victims of the future.

[1]See also `https://iamcybersafe.org/`

6. ACKNOWLEDGMENTS

We would like to thank Brinda Badarinath Hampiholi, Joey de Vries, Lorena Montoya, and Jan-Willem Bullée for their valuable advice and feedback. We would also like to thank the reviewers for their constructive feedback and shepherd Elizabeth Stobert for her helpful comments.

7. REFERENCES

[1] A. Alnajim and M. Munro. An Evaluation of Users' Anti-Phishing Knowledge Retention. In *International Conference on Information Management and Engineering*, ICIME '09, pages 210–214. IEEE, 2009.

[2] I. M. Alseadoon. *The Impact of Users' Characteristics on Their Ability to Detect Phishing Emails*. Phd thesis, Queensland University of Technology, 2014.

[3] Anti-Phishing Working Group. Phishing activity trends report, 3rd quartile 2014, 2014. `http://docs.apwg.org/reports/apwg_trends_report_q3_2014.pdf`.

[4] Anti-Phishing Working Group. Phishing activity trends report, 4th quartile 2014, 2015. `http://docs.apwg.org/reports/apwg_trends_report_q4_2014.pdf`.

[5] M. Blythe, H. Petrie, and J. A. Clark. F for fake: Four Studies on How We Fall for Phish. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 3469–2478, New York, NY, USA, 2011. ACM.

[6] d. boyd. *It's Complicated: The Social Lives of Networked Teens*. Yale University Press, 2014.

[7] C. Brady. Security Awareness for Children. Technical Report RHUL–MA–2010–05, Royal Holloway, London, 2010.

[8] J.-W. Bullee, L. Montoya, M. Junger, and P. Hartel. Telephone-based social engineering attacks: An experiment testing the success and time decay of an intervention. In *Singapore Cyber Security R&D Conference, SG-CRC 2015*, pages 1–6, Singapore, 2016. IOS Press.

[9] J. K. Burgoon and T. R. Levine. Advances in deception detection. In S. W. Smith and S. R. Wilson, editors, *New directions in interpersonal communication research*, pages 201–220. Sage, 2010.

[10] S. L. Calvert. Children as consumers: Advertising and marketing. *Future of Children*, 18(1):205–234, 2008.

[11] J. Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992.

[12] L. E. Cohen and M. Felson. Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44(4):588–608, 1979.

[13] R. Dhamija, J. D. Tygar, and M. Hearst. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, page 581, New York, New York, USA, 2006. ACM Press.

[14] A. Domínguez, J. Saenz-de Navarrete, L. De-Marcos, L. Fernández-Sanz, C. Pagés, and J.-J. Martínez-Herráiz. Gamifying learning experiences: Practical implications and outcomes. *Computers & Education*, 63:380–392, 2013.

[15] L. Haddon and S. Livingstone. EU Kids Online: national perspectives. Technical report, The London School of Economics and Political Science, 2012.

[16] V. Hauch, S. L. Sporer, S. W. Michael, and C. a. Meissner. Does Training Improve the Detection of Deception? A Meta-Analysis. *Communication Research*, pages 1–61, 2014.

[17] J. Hong. The state of phishing attacks. *Communications of the ACM*, 55(1):74–81, 2012.

[18] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Communications of the ACM*, 50(10):94–100, 2007.

[19] D. Kahneman. *Thinking, Fast and Slow*. Penguin Books UK, 2012.

[20] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M. A. Blair, and T. Pham. School of phish: A Real-World Evaluation of Anti-Phishing Training. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, New York, New York, USA, 2009. ACM Press.

[21] P. Kumaraguru, Y. Rhee, S. Sheng, S. Hasan, A. Acquisti, L. F. Cranor, and J. Hong. Getting users to pay attention to anti-phishing education: Evaluation of retention and transfer. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, eCrime '07, pages 70–81, New York, NY, USA, 2007. ACM.

[22] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong. Lessons from a real world evaluation of anti-phishing training. In *2008 eCrime Researchers Summit*, pages 1–12. IEEE, 2008.

[23] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology*, 10(2):1–31, 2010.

[24] A. Lenhart. Teens, Social Media and Technology Overview 2015. Technical report, Pew Research Center, 2015.

[25] E. R. Leukfeldt. Phishing for Suitable Targets in The Netherlands: Routine Activity Theory and Phishing Victimization. *Cyberpsychology, Behavior, and Social Networking*, 17(8):551–555, 2014.

[26] T. R. Levine, H. S. Park, and S. a. McCornack. Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communication Monographs*, 66(2):125–144, 1999.

[27] C. B. Mayhorn and P. G. Nyeste. Training users to counteract phishing. In *Proceedings of the Human Factors and Ergonomics Society*, volume 41, pages 1956–1960, 2012.

[28] K. C. Montgomery, J. Chester, S. A. Grier, and L. Dorfman. The New Threat of Digital Marketing. *Pediatric Clinics of North America*, 59(3):659–675, 2012.

[29] K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, and C. Jerram. The design of phishing studies: Challenges for researchers. *Computers & Security*, 52:194–206, 2015.

[30] S. Purkait. Phishing counter measures and their effectiveness – literature review. *Information Management & Computer Security*, 20(5):382–420, 2012.

[31] E. Rader, R. Wash, and B. Brooks. Stories As Informal Lessons About Security. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, pages 6:1–6:17, New York, NY, USA, 2012. ACM.

[32] S. A. Robila and J. W. Ragucci. Don't be a phish: Steps in User Education. In *Proceedings of the 11th annual SIGCSE conference on Innovation and technology in computer science education*, ITICSE '06, pages 237–241, New York, NY, USA, 2006. ACM.

[33] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. S. Downs. Who falls for phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 373–382, New York, New York, USA, 2010. ACM Press.

[34] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge. Anti-Phishing Phil. In *Proceedings of the 3rd symposium on Usable privacy and security*, SOUPS '07, pages 88–99, New York, New York, USA, 2007. ACM Press.

[35] P. Slovic. Risk-Taking in Children: Age and Sex Differences. *Child Development*, 37(1), 1966.

[36] A. Tsow and M. Jakobsson. Deceit and deception: A large user study of phishing. Technical Report TR649, Indiana University, 2007.

[37] Veilig Bankieren (Dutch Banking Association). Nepmail, daar trapt u niet in, 2011. TV commercial (Dutch): `http://youtu.be/VcbHo0EOtkA`.