NoPhish App Evaluation: Lab and Retention Study

Gamze Canova, Melanie Volkamer, Clemens Bergmann and Benjamin Reinheimer
Technische Universität Darmstadt
name.surname@cased.de

***Abstract*—Phishing is a prevalent issue of today's Internet. Previous approaches to counter phishing do not draw on a crucial factor to combat the threat - the users themselves. We believe user education about the dangers of the Internet is a further key strategy to combat phishing. For this reason, we developed an Android app, a game called –*NoPhish*–, which educates the user in the detection of phishing URLs. It is crucial to evaluate *NoPhish* with respect to its effectiveness and the users' knowledge retention. Therefore, we conducted a lab study as well as a retention study (five months later). The outcomes of the studies show that *NoPhish* helps users make better decisions with regard to the legitimacy of URLs immediately after playing *NoPhish* as well as after some time has passed. The focus of this paper is on the description and the evaluation of both studies. This includes findings regarding those types of URLs that are most difficult to decide on as well as ideas to further improve *NoPhish*.**

I. Introduction

Scammers discover the Internet as a convenient place for their criminal activities. They send Internet users spoofed messages which link to fraudulent but legitimate looking websites. These websites prompt visitors to enter confidential data such as login or banking credentials. This kind of Internet fraud is referred to as phishing. There exist multiple technical approaches to counter phishing, e.g. identifying phishing messages resp. websites (URLs) and adding them to blacklists [1], [2], [3], [4]. Generally, in case a message or URL is blacklisted users are warned e.g. by their email clients or Web browsers. Yet, blacklists cannot guarantee sufficient protection since it takes some time to detect phishing messages/URLs and update the blacklists (according to [5] the average up-time of phishing websites was 32 hours and 32 minutes in the first half of 2014). In this time span users need to be self-reliant and protect themselves against phishing. Moreover, due to the lack of knowledge about phishing and its consequences such warnings (like security warnings in general) are likely to be ignored [6], [7].

For these reasons, a complementary approach is required: user education that makes people aware of phishing, of being a potential victim, and that explains them how to detect phishing URLs (in general and particularly in mobile Web browsers). In particular, users need to learn how to spot URL spoofing tricks phishers exploit to deceive their victims and lure them into disclosing their information on a phishing website. We developed an Android app, a game called –*NoPhish*–, to teach exactly these issues with the aid of well known learning theories such as practice and repetition. The underlying concepts were proposed and discussed in [8].

The aim of this paper is to report about the findings from our effort to evaluate *NoPhish* with respect to its effectiveness and usability. For the evaluation, we conducted a lab study in which we first tested participants' ability to identify phishing websites, then they played *NoPhish* and were tested immediately afterwards again. In addition – and different from earlier research on security education – we conducted an online study five months later. We conducted a retention study because we are convinced that the retention of knowledge is the most crucial part of security education. People do not need to apply their security knowledge frequently, but only in the case of potentially being attacked (in this case receiving a phishing message or visiting a phishing website). In addition, we could show that participants learning with *NoPhish* were significantly better in making proper decisions regarding the legitimacy of URLs both immediately after using the app and after some time has passed. However, we also noticed that a letter swapping based URL spoofing trick (e.g. mircosoft.com) would still be rather successful. Thus, it remains important for companies to search actively for such domains and either make sure they are blacklisted or redirected to the proper domain (as it is actually the case for mircosoft.com). We also found that some participants who were able to detect phishing URLs that contained the brand name in the subdomain (amazon.shopping.com) immediately after playing the game, missed them in the retention study. Therefore, we conclude that it is important to emphasize more on this type of trick in a future version of *NoPhish*.

II. *NoPhish* - Game Design

Before elaborating on our studies to evaluate *NoPhish* we briefly describe the game itself. We applied several learning principles [9], such as exercise, repetition, and direct feedback to optimize learning performance. Gamification elements like lives and levels[1] were also implemented to increase motivation. We followed a user-centered design [10] and involved potential users in the development process. *NoPhish* entails an awareness part and the actual gaming part which are described in the following subsections.

A. Awareness Part

Users are made aware of how simple it is to spoof messages by enabling them to send themselves an email from a sender

USEC '15, 8 February 2015, San Diego, CA, USA

http://dx.doi.org/10.14722/usec.2015.23xxx

[1] http://badgeville.com/wiki/Game_Mechanics Accessed November 26 2014

address *they* provide and with content they provide. Note, this email contains – by default – an explanation that it was sent using the *NoPhish* app to avoid misuse of this functionality. Users are made aware that they should not trust links in messages by providing them with an email that asks them to click on a link displaying 'www.google.com' while clicking on the link leads them back to the app. Finally, the users are made aware that a legitimate looking website does not necessarily mean that it is legitimate by showing them twice the same website (body) but with different URLs in the address bar.

B. Gaming Part

The gaming part is split into eight levels with increasing difficulty. Each level consists of two parts: an introductory block and the actual exercise. The exercise is designed in a playful manner, i.e. users start with three lives, represented by hearts, and can collect points for correct answers and lose points and lives for wrong ones. Users receive direct feedback on their decisions. If the given answer is correct the users are rewarded by gaining points and a smiley face. If the answer is wrong the users lose points and a life. The users are immediately told why their answer was wrong. The next level is achieved if and only if a predefined amount of phishing and legitimate URLs has correctly been identified. The current level also contains exercises from previous levels to repeat those again and again. Figure 1 depicts exemplary screenshots for the introductory and the exercise part. In the following, the learning content of each level is summarized.

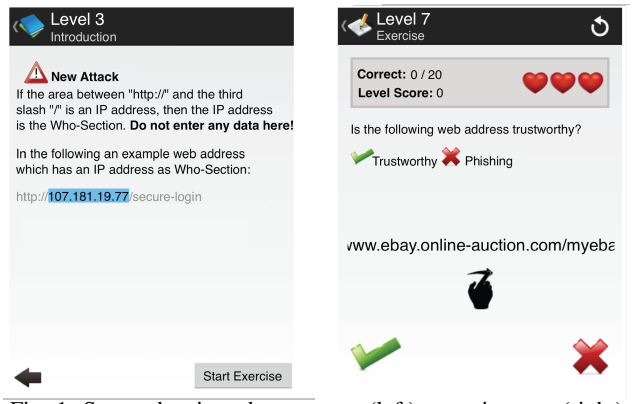

Fig. 1: Screenshot introductory part (left), exercise part (right)

Level 1 - Structure of a URL: Phishers apply different URL spoofing tricks to deceive their victims and lure them into disclosing their data. In this level people achieve the capability of parsing a URL properly as this is a precondition to identify different URL spoofing tricks. Especially, people learn to identify the domain (first- and second-level) in a given URL. As the use of technical terms is avoided, this part of the URL is referred to as *Who-Section*. During the exercise, the user is asked to tap on the *Who-Section*.

Levels 2-8 - URL Spoofing Tricks: In the introductory parts of levels 2-8 existing URL spoofing tricks (based on literature such as [11], [12] and the analysis of PhishTank URLs) are addressed while starting with simple ones that are often applied according to [12] and continuing with more difficult ones:

- Level 2: subdomain tricks with very obvious domain names (e.g. amazon.phishing.com)
- Level 3: IP address tricks (e.g. 5.178.64.164/test)
- Level 4: subdomain tricks with random domain names (e.g. paypal.mayponyfarm.com/)
- Level 5: subdomain tricks with trustworthy sounding or related domain names (e.g. amazon.shopping.com or amazon.secure-login.com)
- Level 6: tricks where the domain name contains some form of the targeted brand name, i.e. either introducing typos (e.g. twpitter.com), swapping letters (e.g. mircosoft.com) or using similar/deceptive domains (e.g. facebook-login.com)
- Level 7: tricks where a character of the domain name is substituted by another (or several) similar looking character(s) (e.g. arnazon.com or paypa1.com)
- Level 8: tricks where the host/brand name appears in the path part of the URL (e.g. auction.org/www.ebay.com)

In the exercise part of the corresponding levels, the user is asked to decide whether the displayed URL a is legitimate or a phishing URL. Every time the user correctly identifies a phishing URL, *NoPhish* asks the user to tap on the *Who-Section*. This way, we aspire to ensure that the user understood where to look at and did not just guess the answer. If the user cannot identify the *Who-Section* the answer to this exercise is considered wrong. Whenever the user gives a wrong answer he receives direct feedback on why the given answer is wrong. The displayed URLs are generated on the fly based on a predefined set of legitimate domains, subdomains, and paths. These were selected from the top Alexa domains for Germany[2]. Spoofing tricks are then applied to these legitimate URLs to get the phishing URLs whenever required. Note that the selection of URLs in *NoPhish* is randomized. For more details about *NoPhish* we refer to [8][3].

III. Lab Study

We conducted a lab and an online retention study to investigate the effectiveness and usability of *NoPhish*. This section describes the hypotheses, gives insights to our participant recruitment, compensation and ethics, explains the design of the lab study and states the results of the lab study. The next section then deals with the results of the online retention study.

A. Hypotheses

We formulated the following hypotheses to evaluate the effectiveness and usability of *NoPhish*:

Hypothesis 1 - Correct Answers: After playing *NoPhish*, the participants give significantly more correct answers when deciding whether or not a website (i.e. the legitimate content and either the legitimate or a spoofed URL in the address bar) is a phishing website than before playing *NoPhish*.

[2]The app is currently only available in German. Both studies were run with Germans in German and screenshots were translated for this paper.

[3]Note, [8] has been published after having analyzed the lab study. While we do not report there about the lab study, we present an improved version of *NoPhish* in [8] – improved based on the results from the lab study. Correspondingly, the levels slightly differ in [8] from what is described here.

Hypothesis 2 - URL Based Decision: After playing *NoPhish*, the participants primarily base their decision whether or not a website is a phishing website significantly more often on the URL.

Hypothesis 3 - URL Comprehension: After playing *NoPhish* the participants primarily base their decision whether or not a website is a phishing website significantly more often on the domain of a URL.

Hypothesis 4 - Good Usability: The usability of *NoPhish* is above average. An SUS score higher than 68 can be considered above average usability.

B. Participant Recruitment, Compensation and Ethics

We used several channels to reach potential participants including flyers distributed in town and online social networks. Those we reached were asked to further advertise the study to get more participants via a snowball approach. Note, when advertising the study, we did not mention the specific topic of phishing in advance because we did not want potential participants to read up about it before the lab study.

People participated in groups of four to five people. The participant who performed best was awarded with a "Golden Anti-Phish Certificate", all other participants received a "Silver Anti-Phish Certificate". For each group a gift certificate was raffled. To express our appreciation to each participant we offered cookies and other kinds of sweets. The participants were informed that there will be a further gift-certificate raffle for the retention study.

The requirements for research that involves humans are set by the university's ethics commission[4]. These requirements are met in our study: at the beginning of the study the participants were explained about the purpose of the study and that they were not obliged to finish it. If they wanted to leave the study before the end, however, they could not be included to the gift certificate raffle. The participants were handed out consent forms with experimental guidelines and with information about the data stored anonymously.

C. Study Design

For the lab study we chose a within-subject design, i.e. a "before and after *NoPhish*" study with the same group of people. People participated in the lab study in groups of four to five people because we wanted to add a motivational game element: the participant who performed best was awarded with a "Golden Anti-Phish Certificate". In the following, we dwell on the different steps of the lab study.

(1) Informed Consent: Before starting the lab study the participants were asked to sign the informed consent form.

(2) Website-Survey Before: In this part of the lab study, each participant got 16[5] printed screenshots of websites, all with legitimate content but eight with phishing URLs and eight with legitimate URLs. The screenshots had been taken with the standard browser of an Android tablet instead of a smartphone due to the small size of a smartphone. When choosing a phishing URL we took care that each spoofing trick (cf. subsection II-B) was represented at least once in this survey. Participants were asked to decide whether or not they would enter confidential data on the shown website. Additionally, for each screenshot they were asked to encircle the part of the screenshot which was the primary reason for their decision. Furthermore, participants were asked to indicate how confident they were about each answer on a five-point Likert scale. Finally, they were asked for each brand used in a screenshot whether or not they knew the brand and had an account there.

(3) Play NoPhish: At this stage, participants got smartphones and had 30 minutes to play *NoPhish*. While playing, the participants were handed out a slip of paper for taking notes. Note that we did not ask the participants to write down anything specific. Afterwards, we collected the smartphones and noted down the achieved points for each level.

(4) Website-Survey After: After playing *NoPhish*, the participants got a second survey. This survey contained all screenshots of the previous survey. Moreover, it contained eight new website screenshots, four with a phishing URL and the other four with the respective legitimate URL.

(5) General-Survey After: In this step, participants were asked to complete some demographic questions. This form also contained questions to compute the System Usability Scale (SUS) [13] as well as questions regarding the participants' impression and opinion of *NoPhish*.

(6) Certificates and Debriefing: Once they had completed the final survey, we thanked the participants and awarded the "Golden Anti-Phish Certificate" and "Silver Anti-Phish Certificates". Next, the gift certificate was raffled. Finally, there was an optional debriefing, where the participants could ask questions or provide their remarks in person.

D. Results for Hypotheses

In total 23 participants attended the lab study[6]. We report here only about those 19 participants who additionally took part in the follow-up retention study.

Demographics: We had six male and 13 female participants. The age of our participants ranged from 20 to 36. Their field of work resp. studies is quite diverse. Three of the 19 participants work/study in the area of electrical engineering or IT. The rest of them work or study in other fields such as economics, finances or architecture.

Measurements: The hypotheses were tested with the aid of the participants' answers and markings. In particular, we tested *hypothesis 1* with the aid of the participants' answers whether or not they would provide their information on a specific website (we considered the percentage of correct answers for each participant). An answer was considered correct if a phishing website was correctly identified as phishing (i.e. the participant would not provide any data) and if a legitimate website was correctly identified as legitimate. All other answers were considered wrong. As we also asked the participants to mark the source of their decision we could test *hypothesis 2* correspondingly. Each marking involving any part of the URL

[4]http://www.intern.tu-darmstadt.de/gremien/ethikkommisson/index.en.jsp

[5]Note, all participants got the same screenshots.

[6]Note, in one group one participant did not show up.

was counted as a URL marking (rather than e.g. content markings). Note that favicons, padlocks or the Web browser tab were not coded as URL markings. Markings involving the domain or parts of the domain (e.g. the position of a typo) were counted as domain markings and used for *hypothesis 3*. Finally, we calculated the usability score of *NoPhish* with the aid of the SUS [13]. An SUS score higher than 68 can be considered above average usability.

Hypotheses 1-3 were tested with the two-tailed Wilcoxon signed-rank test [14] comparing answers from the survey before playing *NoPhish* and answers from the survey after playing *NoPhish*. Table I summarizes the ranks of each hypothesis which we discuss in detail in the following paragraphs. Note, H1a in Table I addresses possible recognition effects relevant for H1. H1a only considers new screenshots from the after survey and compares them against the ones in the survey before.

TABLE I: Ranks of hypotheses 1-3

Hypothesis	H1	H1a	H2	H3
Negative rank	0	0	0	0
Positive rank	19	19	19	18
Binding	0	0	0	1
Total	19	19	19	19

Hypothesis 1: Figure 2 shows the results of both surveys with respect to correct answers. The majority of the participants identified more URLs correctly after playing *NoPhish* than before. While most participants correctly identified eight to ten out of 16 (50-64%) websites before they played *NoPhish*, almost everyone gave correct answers to at least 22 out of 24 (90% and higher) websites afterwards.

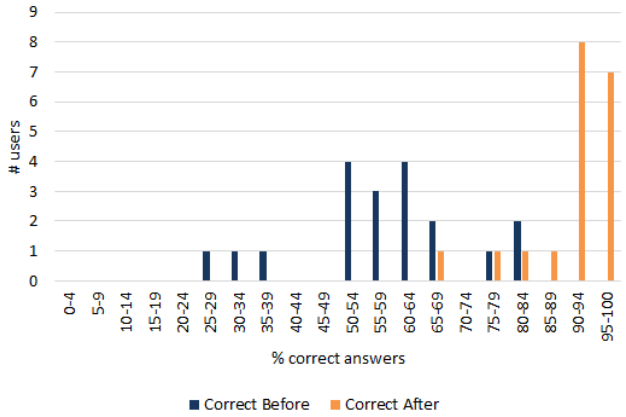

Fig. 2: Correct answers before and after *NoPhish*

Table I shows that *hypothesis 1* only reveals positive ranks, i.e. each participant gave more correct answers (in percent) compared to the survey before. Applying the two-tailed Wilcoxon signed-rank test we found out that the group survey before and the group survey after differ significantly with a p-value of 0.001 in their median of correctly answered questions. The test shows the survey after group gave significantly more correct answers compared to the survey before group. Thus, this result supports our *hypothesis 1*.

One could argue that this increase in correct answers is based on the fact that the examples are mainly the same in the survey after, i.e. the reason for the participants' improved performance is based on recognition effects. Thus, we additionally applied the test while comparing the percentage of correct answers wrt. new URLs of the survey after with the ones of the survey before. Column H1a of Table I contains the ranks for this test. The negative rank of zero indicates that no participant performed worse considering the new screenshots during the survey after compared to the screenshots of the survey before. All participants performed better (positive rank of 19). Applying the two-tailed Wilcoxon signed-rank test indicates that the survey after group with new screenshots differs significantly from the survey before group with a p-value of 0.001. The survey after group with new screenshots gives significantly more correct answers than the survey before group. Thus, recognition effects can be neglected.

Hypothesis 2: Figure 3 shows how many participants marked URLs as their main source of decision. Most of the participants already based most of their decisions on the URL before playing *NoPhish*. E.g. seven participants based their decision on the URL in 90-94% of the screenshots and another five participants based their decision on the URL in even 95-100% of the cases. In the before survey, we identified 35 markings for the content of the website or the padlock (out of 303 markings). After playing *NoPhish* the website content was marked only once, none marked the padlock[7].

Comparing the before and after survey with respect to URL markings we can observe that five participants (approx. 26%) always marked the URL in the survey before playing *NoPhish* (note that Figure 3 indicates the span of 95-100%, but in fact all five participants always marked the URL, i.e. in 100% of the cases). After playing, 15 participants (approx. 79%) always based their decision on the URLs, one participant based his decision on the URL only 55-59% of the cases and the remaining three participants made an exception only once or twice. Table I shows that *hypothesis 2* only reveals positive ranks, i.e. each participant marked the URL more often after playing *NoPhish* compared to the survey before. Applying the two-tailed Wilcoxon signed-rank test we found that the group survey before and the group survey after differ significantly with a p-value of 0.001 in their median of marked URLs. Thus, the survey after group marked the URL significantly more often than the survey before group. This result supports *H2*.

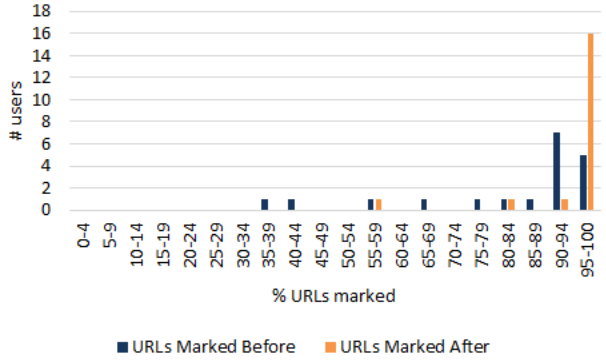

Fig. 3: URLs marked before and after *NoPhish*

Hypothesis 3: This hypothesis deals with the comprehen-

[7]Note that we focused on the detection of phishing URLs rather than SSL.

sion of the URL. Figure 4 indicates the identified marked areas in the URL before and after playing *NoPhish*. This figure indicates a shift of the participants' focus. Afterwards domains or parts of domains were marked in the majority of the cases. Table I shows that *hypothesis 3* mostly reveals positive ranks, i.e. 18 of the 19 participant marked the domain or a part of the domain (e.g. the part with a typo) more often after playing *NoPhish* compared to the survey before. There is only one participant where the amount of markings was the same for the survey before as well as survey after (binding = 1). Applying the two-tailed Wilcoxon signed-rank test we found that the group survey before and the group survey after differ significantly with a p-value of 0.001 in their median of marked domains or marked parts of the domains. Thus, the survey after group marked the domain or parts of the domain significantly more often compared to the survey before group. This result supports our *hypothesis 3*.

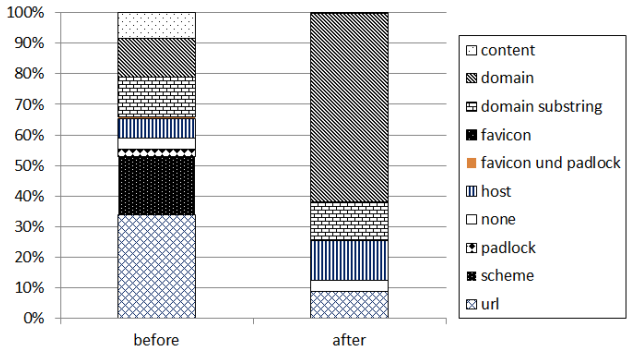

Fig. 4: Marked areas before and after *NoPhish*

Hypothesis 4: This hypothesis states that the usability of *NoPhish* is above average, i.e. has a SUS score above 68. Based on the answers in the SUS section, we computed a SUS score of 83.7. This is considerably higher than 68 which highlights the satisfaction of our participants with our app.

E. Further Findings

In addition to the elaboration on our hypotheses, we conducted some further analyses regarding participants' confidence and their opinion about *NoPhish*.

Confidence: For each screenshot we asked the participants how confident they were about their answers on both website-surveys with a five-point Likert scale. Before playing *NoPhish* only four of 19 participants stated a median confidence of five. Afterwards, there were 15 out of 19 (thus eleven additional) participants who indicated five in median. We ran a statistical test to determine the statistical significance. We applied a two-tailed Wilcoxon signed-rank test again. We found out that the answers in the before and in the after survey differ significantly with a p-value of 0.001 in their median confidence. Thus, the participants after playing *NoPhish* had a significantly higher median confidence than before playing it.

Participant opinions: The general-survey after contained statements which the participants had to assess with the aid of a five-point Likert scale. The corresponding median answers are:

- 5 for '*NoPhish* helped me identify phishing websites in future'
- 3.5 for 'I was motivated by the spoofed email to continue playing the app'
- 3.5 for 'The amount of texts was appropriate'
- 5 for 'The text was easy to understand'

IV. Retention Study

We conducted a retention study (about five months later) to determine how much participants are able to retain.

A. Hypotheses

To test and evaluate how well participants could retain the knowledge they gained with *NoPhish*, we want to test the following hypotheses wrt. correct answers.

Hypothesis 5 - Website-survey before (lab study) vs. website-survey retention: In the retention study the participants give significantly more correct answers when deciding whether a website (i.e. the legitimate content and either the legitimate or a spoofed URL in the address bar) is a phishing website or not than *before* playing *NoPhish*.

Hypothesis 6 - Website-survey after (lab study) vs. website-survey retention: In the retention study the participants *do not* give significantly fewer correct answers when deciding whether a website is a phishing website or not than directly *after* playing *NoPhish*.

B. Study Design

The retention study was realized as an online survey. Every participant who had participated in the lab study received an email with the corresponding link. The online survey consisted of mainly two parts which were:

(1) General Questions: The participants were asked general questions regarding phishing, e.g. whether they received a phishing email in the last few months. Furthermore, they were asked whether *NoPhish* helped them identify phishing attacks.

(2) Phishing Survey: This part is similar to the survey parts from the lab study: The participants saw screenshots of websites (containing legitimate content and either the legitimate or a spoofed URL in the address bar) in their Web browser. In this survey, all 24 examples of the lab study are included. Moreover, it contains ten new website screenshots of which five represent phishing and the other five legitimate URLs. Thus, the participants were confronted with 34 screenshots in total. The participants were asked to reason their answers for 16 websites, eight legitimate and eight phishing websites. They were again asked to indicate their confidence about each answer they gave.

C. Results for Hypotheses

In total 19 participants answered the follow-up retention survey and are as such included in the following findings.

Measurements: The *hypotheses 5 and 6* were tested similarly to *hypothesis 1* with the aid of the answers from the lab and the retention study.

Hypotheses 5 and 6 were also tested with a two-tailed Wilcoxon signed-rank test with the 19 participants from the

website-survey before resp. after (playing *NoPhish*) and the same participants from the website-survey retention. Table II shows the scores for all participants for each study - before, after, and retention. Table III summarizes the ranks of the hypotheses which we discuss in detail in the following paragraphs. Note, H5a in Table III considers a special case of H5.

TABLE II: Score per participant: before - after - retention

Participant	% Before	% After	% Retention
1	56.25	95.83	67.65
2	62.5	87.5	91.18
3	62.5	95.83	73.53
4	81.25	100	97.06
5	62.5	91.67	79.41
6	50	95.83	76.47
7	68.75	91.67	97.06
8	56.25	91.67	82.35
9	81.25	91.67	73.53
10	50	79.17	76.47
11	62.5	91.67	100
12	50	91.67	82.35
13	50	66.67	61.76
14	75	95.83	82.35
15	68.75	95.83	76.47
16	31.25	83.33	70.59
17	37.5	91.67	94.12
18	25	91.67	73.53
19	56.25	95.83	100
Average	57.24	90.79	81.89

TABLE III: Ranks of hypotheses 5 and 6

Hypothesis	H5	H5a	H6
Negative rank	1	2	15
Positive rank	18	17	4
Binding	0	0	0
Total	19	19	19

Hypothesis 5: Figure 5 depicts the distribution of correct answers for each of the surveys. Participants perform still better than they did for the website-survey before.

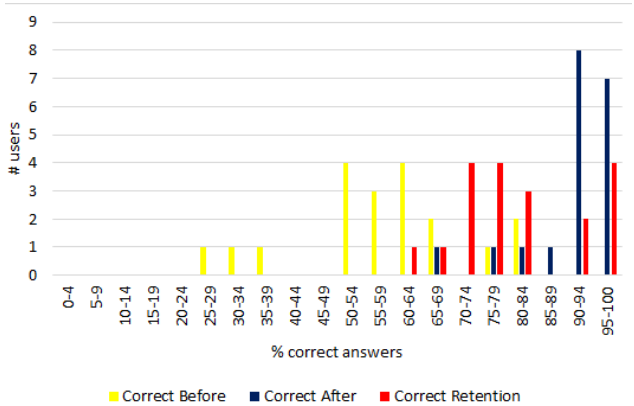

Fig. 5: Correct answers lab and retention study

Table III shows that *hypothesis 5* mostly reveals positive ranks, i.e. 18 participant gave more correct answers for the website-survey retention (five months after playing *NoPhish*) compared to the website-survey before playing *NoPhish*. Only one participant performed worse. Applying the two-tailed Wilcoxon signed-rank test we found out that the group survey before and the group survey retention differ significantly with a p-value of 0.001 in their median of correctly answered questions. Participants gave significantly more correct answers in the retention survey compared to the survey before they played *NoPhish*. Thus, this result supports our *hypothesis 5*. Again, we additionally applied the test while only considering new URLs for the website-survey retention (cf. H5a in Table III). The results show even if only new URLs are considered the group survey before and the group survey retention still differ significantly with a p-value of 0.001 in their median of correctly answered questions. The participants also better distinguish *new* phishing URLs from legitimate ones after approximately five months compared to before they played *NoPhish*.

Hypothesis 6: Figure 5 depicts the distribution of correct answers for each of the surveys. Comparing the performance in the retention study with the performance directly after playing the app one can see that the participants' performance has decreased. This degradation is also revealed in the ranks of Table III. 15 participants have a negative rank, i.e. they perform worse in the website-survey retention compared to the website-survey after. However, there are four participants who perform better in the website-survey retention. Applying the two-tailed Wilcoxon signed-rank test we found out that the group survey after and the group survey retention differ significantly with a p-value of 0.005 in their median of correctly answered questions. The survey after group gave significantly more correct answers compared to the survey retention group. Thus, this result leads to the rejection of *hypothesis 6*.

D. Further Findings

This subsection deals with results from the additional questions. Regarding the first two the results are:

- Eleven of 19 participants stated they detected a phishing attack ever since.
- All of them except for one participant stated that the *NoPhish* app played an important or very important role for the ability to detect the phishing attack.

In all website-surveys we asked the participants how confident they were about each of their answer on a five-point Likert scale. Figure 6 depicts the medians of the participants' confidence for all surveys (before, after, retention). Before playing *NoPhish*, four participants stated (in median) a confidence level of five. Immediately after playing *NoPhish*, there were 15 participants who stated five in median. One can clearly see that the confidence decreased compared to the median confidence indicated in the survey after. After five months only 13 (still the majority) selected a confidence level of five in median.

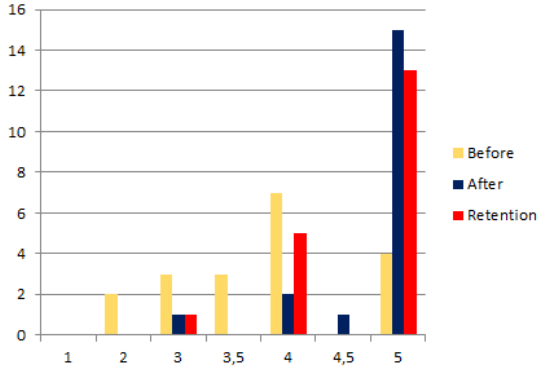

Fig. 6: Median confidence for all studies

V. Discussion

This section discusses the results of the lab and retention study. After discussing the results of the hypotheses, we provide further insights into results for individual URLs to deduce possible improvements for the app. Finally, we dwell on the limitations of our study.

A. Discussion of Hypotheses

The results of both, the lab as well as the retention study are very promising. However, there was a significant decrease in performance when comparing the website-survey after with the website-survey retention (H6). As this is not what we expected, we discuss potential reasons that may have caused this result. *1) Regression towards the mean:* After playing *NoPhish* the average score the participants achieved was about 90% (with 79% of the participants scoring 90% and higher). About 37% of the participants even achieved about 95% or higher. In case one or more persons' performance decrease it is very difficult to balance this degrading with much better results than 90% as it is not possible to achieve more than 100%. *2) The exponential nature of forgetting:* Ebbinghaus, e.g. [15] found out that even after 20 minutes only 60% of the newly gained knowledge can be recalled. After one day only 34% and after 6 days only 23% can be recalled. As the experiment conducted by Ebbinghaus was based on a sequence of syllables the numbers might of course look different for different situations, however, they give an indication of how fast a human forgets learnt material. *3) Spacing effects:* It is easier to learn and remember learnt material when the material is studied a couple of times spaced over a longer period of time compared to studying new material repeatedly in a short time span [15]. Thus, the retention of the participants might have not decreased significantly if they were asked to repeat the content in a spaced manner over a longer time period.

B. Analysis and Discussion of Results for Individual URLs

We start with the various phishing URLs and continue with the legitimate ones. Figure 7 depicts the mistakes (in %) for each website-survey and each spoofing trick covered in the surveys. It distinguishes between those who completed the level and those who did not. Note that level 1 and 2 are not covered in the surveys: level 1 deals with the identification of the domain and level 2 covered very obvious URL spoofing tricks such as "amazon.phishing.com" which is why we did not include it to the surveys.

IP address trick: Before playing *NoPhish* most participants already recognized the IP address trick. After playing the app and also in the retention study no participant fell for it.

Subdomain trick: All other URL spoofing trick types seem to be a problem before playing *NoPhish* in some way, especially the subdomain trick. Directly after playing *NoPhish* the participants' performance increased (they made fewer mistakes) for almost every URL spoofing trick including the subdomain trick. However, after five months the subdomain seems to be a problem again. Further analyzing the participants who fell for the subdomain trick we found out that in the website-survey before twelve participants fell at least for one subdomain trick. In the website-survey after there were only two participants who fell for at least one subdomain trick, both fell for at least one subdomain trick in the website-survey before already. One of these two participants completed only level 3, meaning that he only learnt about the very obvious subdomain trick like "amazon.phishing.com" and the IP address trick. In the website-survey retention eight participants fell for at least one subdomain trick. Both participants, who already fell for subdomain trick(s) in the website-survey before and after are represented there as well. Only one of these remaining six participants had not made any mistakes for the website-survey before or the website-survey after with respect to subdomain tricks. All other five participants fell for at least one subdomain trick in the website-survey before.

Letter swapping trick: An exception where participants did not do better after *NoPhish* as well as in the retention study is the letter swapping trick. The trick with the swapped letters is a tricky problem as it is very difficult to detect it unless you read the URL *very* carefully, even then it is likely to remain undetected. An explanation for this phenomenon might be that the order of letters does not matter because the human mind does not read every single letter, but the word in a whole instead [16], [17]. However, according to [18] scrambling letters comes with a cost which depends on at which position the letters are scrambled. It is worth mentioning that about 1.7% of all maliciously registered *domains* (note we do not refer to the complete URL) contain the targeted brand name in some form [12]. This tactic is not popular because usually brand owners scan Internet zone files for their own brand names. Even if 1.7% sounds not a lot, looking at the absolute value 1.498 it is a reasonable number (not all of them include the letter swapping).

Similar/deceptive domain trick: Analyzing the participants who fell for the similar/deceptive domain tricks we found out that in the website-survey after only one person fell for the "paypal-secure.com" trick. This person had achieved level 7, i.e. learnt about the similar/deceptive domain trick. This person knew the provider PayPal, however, did not have an account there. In the website-survey retention four participants fell for this trick. The participant who fell for this trick in the website-survey after did not fall for it in the website-survey retention, i.e. all four participants are other participants. One person of these four knows PayPal and also has an account there. This participant did not fall for this trick in the website-survey before as well as in the website-survey after. Thus, it is possible the participant has overlooked it (due to lack of concentration). Two participants who fell for this trick neither know PayPal nor have an account at PayPal which means that it is possible that these participants do just not know how the actual domain of PayPal looks like. The fourth participant indicated that he did not know PayPal but had an account there which seems to be a mistake.

Legitimate URLs: Having a closer look at legitimate URLs we can observe the following for the surveys: in the website-survey before in average 4.59 legitimate websites were falsely identified as phishing. As the participants had to assess eight legitimate websites in total we can say that about 57% of the legitimate websites were falsely rejected. It seems like the participants were overcautious. Directly after playing *NoPhish* the participants falsely rejected only 0.84 of twelve (7%) legitimate websites in average. The participants seem to have lost their fear and can distinguish legitimate websites from phishing ones directly after playing the app. Finally, in the

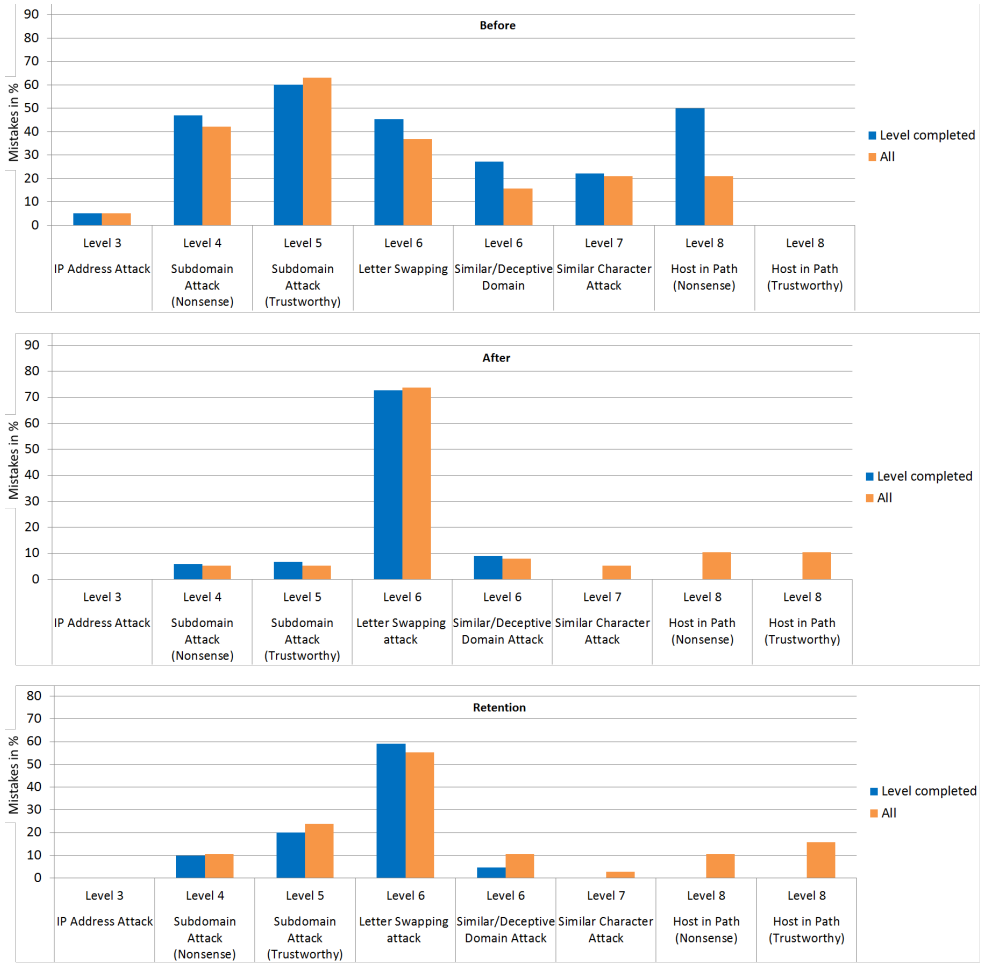

Fig. 7: Wrong answers for specific URL spoofing tricks

website-survey retention 3.47 of 17 (22%) legitimate websites were falsely rejected. After five months the participants got overcautious again, but still not as uncertain and overcautious as in the website-survey before.

We further analyzed the explanations the participants gave to their answers. Their answers indicate that the subdomain in fact seems to be a problem. The participants seem to think there is not supposed to be anything else but the domain in the host part of a URL. Furthermore, they seem to be confused by terms like "secure" occurring in the URL or subdomain. Thus, there seems to be need for improvement in these aspects.

C. Limitations

Study limitations are discussed in this subsection.

Within-Subject Study Design: Within-subject deals better with variability associated with individual differences compared to between-group design, where different groups would be considered who do and do not play *NoPhish*. A major drawback of the within-subject design, however, is the recognition effect. The result of our study, however, shows that recognition effects can be neglected (cf. subsection III-D).

Behavior Change: In the lab study the participants were not in their usual environment. Therefore, they likely behaved differently during the study [6], [19]. An alternative approach would have been to distribute *NoPhish* to several participants and ask them to play it remotely. However, this has two major downsides: First, the participant would have been remote and thus we would have less resp. no control over the conditions. Second, testing the before and after app skills would have been difficult to realize and to ensure a homogeneous process among all participants.

Increased Attention: In both studies, the lab and the retention study, the participants were told that the study dealt with phishing. Additionally, they were explicitly asked to indicate whether the websites were phishing websites or not. Thus, the participants automatically increased their attention towards answering these kinds of questions. Designing an in situ study, where the participants would have been in their usual environment and would have not known about their participation, was not considered because such a design is ethically questionable.

Changed Question: In the lab study the participants were asked "Would you enter sensitive information on this website". In the retention study we asked "Is this website a phishing website". One could argue that this change might mean that the results are not comparable. However, Figure 3 shows that even before playing *NoPhish* participants based their decision

on the URL in most cases already which indicates that they looked for phishing indications and were not motivated by personal reasons not to provide information on a website.

Domain Markings: There is a general issue with the question for the test of H3 in the websites-surveys. In the website-survey before we were not able to clearly ask the participants to mark the domain when it was the reason for their decision because we would have then pointed them towards looking at the URL or even at the domain. This would have had an impact on the results of H2. Since we could not formulate this question clear-cut, a participant might have marked the complete URL even if his decision was based on a part of it only (e.g. the domain). Consequently, we are not able to clearly identify what the participants' main source of decision was in the website-survey before when they marked the entire or several parts of the URL. We were aware of this problem beforehand, but saw no other option than formulating the question in such an open form. After playing *NoPhish* the participants knew that they were expected to mark the domain (due to level 1 and marking of the *Who-Section* in case a phishing URL was detected in *NoPhish*). This can be interpreted as a change of the question even if the question did not change literally.

VI. Improvements for *NoPhish 2.0*

Despite the limitations of our studies we are confident that we obtained a good insight into the effectiveness of *NoPhish*. Here we present what we already changed in the app and what we plan to change in future based on the study results and the feedback we received.

Improvements based on the findings: Retention is crucial. There seems to be some degree of retention, however, it is necessary to repeat the learnt content over a longer time period on a regular basis. In this form retention can and should be optimized. *NoPhish* should be extended in such a way, that it tests its user with a few questions after some time again and again. It is important not to include too many questions in order not to annoy users. This is part of future work.

Subdomain tricks were challenging in the retention study. As in about 60% of the cases phishers exploit the brand name by using it in some form in the URL, either in the subdomain or in the path part [5], it is crucial to teach the users the difference. We plan to extend the corresponding levels and include more such examples in the exercises (while less obvious ones like IP address tricks). Letter swapping was a challenge in both studies. Currently, the letter swapping is explained in one level together with the typo (generalized form of letter swapping) and similar and deceptive domains. Considering the results for letter swapping we plan to separate this URL spoofing trick to an additional level and include more examples. Furthermore, this type of trick could be visualized and explained in more detail, e.g. by means of a complete text containing words and sentences with swapped letters. Such a text would exepmlify how diffiuclt it is to detect such scrabmled lettres.

Improvements based on participants' remarks: Some participants remarked that a couple of URLs referred to services they did not know, even though we only used websites of Alexa's top rankings. Therefore, the current version of *NoPhish*, always provides the information "You want to visit the website of *provider*" in the exercise part before asking whether it is a phishing URL or not. A few participants mentioned issues with the positioning of some buttons which we also improved in the current version of *NoPhish*. In every introductory part of a level we briefly repeat the so far learned parts of a URL (with a graphic) and the different URL spoofing tricks the user has seen until this point. Some of our participants explicitly indicated that our repetitions made them feel more confident and safer. However we received some personal feedback that the repetitions in the introductory parts were too much and partially unnecessary as they just played the previous levels. Thus, we adjusted *NoPhish* in such a way that the screen for the repetition in the introductory block is only displayed when a level is started from the main menu.

VII. Related Work

We start with previous work on anti-phishing education. Afterwards we proceed with related work in the area of security education and awareness in general and their evaluation.

Anti-Phishing Phil [11] is a well-known anti-phishing education game. The main character of the game is Phil. Phishers try to trick Phil into eating their fake worms (phishing URLs). For the reasons why we chose not to follow the approach of Anti-Phishing Phil, we refer to [8]. To evaluate the effectiveness of the game the authors conducted a between-subjects experiment with three training conditions: (1) existing training material, e.g. from eBay or Microsoft, (2) anti-phishing tutorials created based on the game, and (3) the game itself. Each group had to decide on ten websites (in total 20) about their legitimacy before and after the training. The results showed that the participants in the game condition performed better than those in the other two conditions. Sheng further analyzed the effectiveness of Anti-Phishing Phil in an online study and also conducted a retention study for Anti-Phishing Phil [20]. The participants were able to retain most of the gained knowledge. However, the retention study was conducted only seven days after the online study. We evaluate the retention five month after playing *NoPhish*. This shows that it is important to find out when repeating parts of the content is necessary - maybe not after a week but definitively earlier than five months.

Another proposal to educate users in protecting against phishing is embedded learning [21], [22] where simulated phishing emails with links to fake websites or malicious attachments are sent to users. The moment a user falls for a trap he receives a notification informing him that he could have fallen for a real phishing attempt. This notification email includes a link to a website with a training program and hints on how to detect phishing and malicious attachments. This approach was evaluated in [22]. In a first lab study the authors compared its effectiveness with typical email security notices. Their results suggest that the embedded learning approach is more effective. A knowledge retention study was conducted in a second lab study. They report no significant difference between performances directly after the training and in the retention study seven days later. This again shows that it is important to find out when repeating parts of the content is necessary.

Smith et al. [23] propose an awareness raising website covering risks like phishing, spam or pop-ups. They evaluated

the effectiveness. Participants reading the website material seem to perform better in answering the corresponding quizzes than participants who did not read the material. A retention study was not conducted.

VIII. Conclusion and Future Work

We developed an anti-phishing education app –*NoPhish*– which educates the user to detect phishing URLs in a playful manner. We evaluated *NoPhish* and received promising results and also insights to aspects which could be further improved. The participants performed significantly better directly after playing *NoPhish* in all aspects: they gave significantly more correct answers (15 of the 19 participants gave at least 90% correct answers), they based their decision significantly more often on the URL (after playing *NoPhish* 15 participants *always* marked the URL) and they marked the domain or parts of the domain significantly more often. We conducted a retention study after five months because it is a crucial part in security education: people are not confronted with attacks on a daily basis. Thus, they do not apply their knowledge regularly. Participants still performed significantly better than before playing *NoPhish*. However, participants' performance decreased compared to immediately after playing *NoPhish*.

We identified the tricks of letter swapping as the most difficult one (also immediately after playing *NoPhish*). While future versions of *NoPhish* will focus more on this attack, we also recommend that service providers search the web for domain names that match the own one while letters are swapped. In addition, we noticed in the retention study, that subdomains are an issue in both situations: corresponding phishing URLs and legitimate URLs. Corresponding changes are planned to be integrated in *NoPhish* to avoid confusion. Furthermore, we plan to address the spacing effect by extending *NoPhish* in such a way that users are tested on a regular basis over a longer time span with a few examples. In future studies we plan to address people from other age ranges, younger and older than in these studies. We also want to compare *NoPhish* with other training conditions such as Anti-Phishing Phil, reading or video material and design a study which is closer to reality to avoid the focus on security as primary task. Once finalized *NoPhish* will be made available at the Google Play Store and evaluated in particular regarding the necessary retention.

Acknowledgment

This work was supported by CASED and EC SPRIDE.

References

[1] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious urls," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 1245–1254.

[2] P. Prakash, M. Kumar, R. Kompella, and M. Gupta, "Phishnet: Predictive blacklisting to detect phishing attacks," in *INFOCOM, 2010 Proceedings IEEE*, March 2010, pp. 1–5.

[3] S. Marchal, J. François, T. Engel *et al.*, "Proactive discovery of phishing related domain names," in *Research in Attacks, Intrusions, and Defenses*, ser. Lecture Notes in Computer Science, D. Balzarotti, S. Stolfo, and M. Cova, Eds. Springer Berlin Heidelberg, 2012, vol. 7462, pp. 190–209.

[4] Z. Ramzan, "Phishing attacks and countermeasures," in *Handbook of Information and Communication Security*, P. Stavroulakis and M. Stamp, Eds. Springer Berlin Heidelberg, 2010, pp. 433–448.

[5] G. Aaron, R. Rasmussen, and A. Routt, "Global phishing survey: trends and domain name use in 1h2014," *Anti-Phishing Working Group (APWG), Lexington, MA*. [Online]. Available: http://docs.apwg.org/reports/APWG_Global_Phishing_Report_1H_2014.pdf

[6] M. Wu, R. C. Miller, and S. L. Garfinkel, "Do security toolbars actually prevent phishing attacks?" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '06. New York, NY, USA: ACM, 2006, pp. 601–610.

[7] D. Akhawe and A. P. Felt, "Alice in warningland: A large-scale field study of browser security warning effectiveness," in *Presented as part of the 22nd USENIX Security Symposium (USENIX Security 13)*. Washington, D.C.: USENIX, 2013, pp. 257–272.

[8] G. Canova, M. Volkamer, C. Bergmann, and R. Borza, "NoPhish: an anti-phishing education app," in *Security and Trust Management*, ser. Lecture Notes in Computer Science, S. Mauw and C. Jensen, Eds. Springer International Publishing, 2014, vol. 8743, pp. 188–192.

[9] E. Thorndike, "The fundamentals of learning. 1932," *Teachers College Bureau of Publications, New York*, 1932.

[10] C. Abras, D. Maloney-Krichmar, and J. Preece, "User-centered design," *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*, vol. 37, no. 4, pp. 445–56, 2004.

[11] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge, "Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish," in *Proceedings of the 3rd Symposium on Usable Privacy and Security*, ser. SOUPS '07. New York, NY, USA: ACM, 2007, pp. 88–99.

[12] Anti-Phishing Working Group et al., "Phishing activity trends report. 2nd quarter," 2014. [Online]. Available: http://docs.apwg.org/reports/apwg_trends_report_q2_2014.pdf

[13] J. Brooke, "SUS: A Retrospective," *Journal of Usability Studies*, vol. 8, no. 2, pp. 29–40, 2013.

[14] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in Statistics*, ser. Springer Series in Statistics, S. Kotz and N. Johnson, Eds. Springer New York, 1992, pp. 196–202.

[15] H. Ebbinghaus, *Memory: A contribution to experimental psychology*. New York by Teachers College, Columbia University, 1913, no. 3.

[16] G. E. Rawlinson, "The significance of letter position in word recognition." Ph.D. dissertation, University of Nottingham, 1976.

[17] G. Rawlinson, "Reibadailty," *New Scientist*, vol. 162, no. 2188, pp. 55–55, 1999.

[18] K. Rayner, S. J. White, R. L. Johnson, and S. P. Liversedge, "Raeding wrods with jubmled lettres there is a cost," *Psychological science*, vol. 17, no. 3, pp. 192–193, 2006.

[19] S. Egelman, J. King, R. C. Miller, N. Ragouzis, and E. Shehan, "Security user studies: Methodologies and best practices," in *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '07. New York, NY, USA: ACM, 2007, pp. 2833–2836.

[20] X. S. Sheng, "A policy analysis of phishing countermeasures," Ph.D. dissertation, Carnegie Mellon University, 2009. [Online]. Available: http://www.chariotsfire.com/thesis/SteveSheng_Thesis_Final.pdf

[21] K. Jansson and R. von Solms, "Simulating malicious emails to educate end users on-demand," in *Web Society (SWS), 2011 3rd Symposium on*, Oct 2011, pp. 74–80.

[22] P. Kumaraguru, "Phishguru: A system for educating users about semantic attacks," Ph.D. dissertation, Carnegie Mellon University, 2009. [Online]. Available: http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA501765

[23] A. Smith, M. Papadaki, and S. Furnell, "Improving awareness of social engineering attacks," in *Information Assurance and Security Education and Training*, ser. IFIP Advances in Information and Communication Technology, J. Dodge, Ronald C. and L. Futcher, Eds. Springer Berlin Heidelberg, 2013, vol. 406, pp. 249–256.