Phishing, Data-Disclosure and The Cognitive Reflection Test

Ingvar Tjostheim
Norwegian Computing Center
ingvar@nr.no

Abstract

Phishing is a form of online identity theft that aims to steal sensitive information such as online banking passwords and credit card information from users. Data is key for the digital economy, but disclosing personal data online increases vulnerabilities and the likelihood of experiencing negative consequences from disclosure. In this paper, we analyze willingness to share personal data, a preference for an intuitive decision style and susceptibility to phishes. We report the results of three large-scale national studies in Norway that included the cognitive reflection test (CRT) and a choice experiment on willingness to share personal data. With a binary logistic regression method, we analyzed the relationship between the CRT, willingness to share data and demographical variables with susceptibility to phishes as the outcome variable. Our main finding is that the willingness to share personal data and an intuitive thinking style significantly predict the probability of falling for phish. These results are based on three large-scale studies with national populations, in contrast to earlier studies that in most cases relying on student populations, giving them greater validity.

1. Introduction

Phishing is a confidence trick aimed at getting unsuspecting people to give away personal details on the internet so that the perpetrator can make fraudulent use of their credentials [1], [2]. Often, phishing attacks are very sophisticated so that even well-educated and cautions internet users are liable to fall for phishing. Many internet users have a tendency towards privacy-compromising behavior, revealing a divergence between privacy attitudes and actual behavior [3], [4]. According to Nicholson et al. [5] phishing is an example where users are overconfident. Other factors are inattention, optimism biases, lack of rational behavior, limited mental resources and other "biases and heuristics - well known to behavioral researchers" (Acquisti et al. 2017: 32, [6]). Our study targeted a national population by recruiting participants through two professional market research companies. The study included questions about phishing and misuse of personal data and a choice experiment on sharing of personal data. The Cognitive Reflection Test [7] was used as a measure of individuals' ability to suppress intuitive and spontaneous wrong answers in favor of correct answers requiring greater reflection. According to Toplak et al. [8], the CRT has the capacity to function as a unique predictor of performance on a number of heuristics-and-biases tasks.

2. Related work and motivation for the research

Ferreira & Vieira-Marques [9] give an overview of ten years of phishing research based on 605 scientific journal abstracts. They conclude that there is no single solution for the phishing threat and, for future research, call for a "*focus on socio-technical and integrated solutions that can reflect a comprehensive understanding of both human computer interaction and* ***user unique characteristics***" (our emphasis). Addressing this need to assess user unique characteristics was a main motivation for this research.

According to Volkamer et al. [10] and the APWG Internet Policy Committee Global Phishing Survey it takes, on average, 28.75 hours to detect new phishing websites. Users are mostly unprotected from phishing until malicious websites are identified and blocked [11]. To avoid phishing during this period, users have to reflect on whether to go along with what they are being asked to do (for a phish to work), rather than simply complying. This motivates our research into intuitiveness (automatic decision-making behavior) versus reflective problem-solving styles in relation to the tendency to fall for phishing and willingness to share personal data, and why we chose to include a version of the CRT.

2.1. The Cognitive Reflection Test, phishing studies and sharing of personal data

URI: https://hdl.handle.net/10125/80268
978-0-9981331-5-7

The CRT is often thought of as measuring “people’s tendency to answer questions with the first idea that comes to their mind without checking it” (Kahneman, 2011:65, [12]. This has been attributed to a tendency towards “miserly” information processing, to impulsively accept the solution to a problem that involves expending a minimum of cognitive effort [8], [13]. To score highly on the CRT, the respondent needs to reflect on and question their initial intuitive responses [14], [15] and this involves cognitive effort. This corresponds to the personal tendency not to rely on intuition (which is fast), rather than analytical reasoning (which takes longer).

Bialek & Pennycook [16] discuss whether or not the cognitive reflection test is robust to multiple exposures. They suggest that it is and write that “…participants who do poorly on the CRT massively overestimate their performance (i.e., they do not realize they are doing poorly; Pennycook et al., 2017), indicating that intuitive individuals may have a metacognitive disadvantage (see also Mata et al., 2013)” [17].

It could be argued that low scores on the CRT simply reflect low mathematical skill or general cognitive ability. But while these factors may influence their scores somewhat, they do not explain them completely [18], [19], [20], [8], [13]. The CRT aims to cue intuitions that are common across people and lead to potential responses from nearly all test-takers. Differences in scores can then be taken to reflect an individual’s tendency towards reflective versus intuitive thinking. We suggest that the CRT is relevant for phishing research, since in a phishing context a fast and intuitive response style might be expected to correlate with higher vulnerability.

Several studies have used the CRT in relation to phishing susceptibility, though not with national populations. Kumaraguru et al. [21] in a study with 42 students in a lab experiment, found the low CRT score group had a higher probability of clicking on the phishing-no-account e-mails than those in the high CRT score group, 0.39 versus 0.04, respectively. In their study with the classic three-items CRT, a CRT score of 0-1 (all wrong or one correct) was coded as the “low CRT group” and 2-3 (two or all correct) as the “high CRT group.”

Butavicius et al. [22] performed a phishing study with 121 students. These researchers found a significant negative correlation between CRT scores and link safety judgments for spear-phishing ($\rho < -.23$, $p < .014$, $N = 112$) and phishing ($\rho < -.3$, $p < .001$, $N = 112$) emails, but no significant correlation between performance on the CRT and link safety judgments on genuine emails ($\rho < -.01$, $p < .973$, $N = 114$). Petraityte et al. [23] recruited 100 participants consisting of university students, lecturers and staff, and asked them to assess QR-codes. They found that less impulsive people who did not know what the purpose of the test was (those with a higher CRT score) responded better. Participants with higher CRT scores were less likely to click on the URL held inside the fake QR code. Cognitive impulsivity did not reveal any significant difference for the participants who were informed what the study was about. Finally, in a study by Jones et al. [24] with 224 university students and staff, the participants were asked to examine 36 emails (18 legitimate and 18 phishing emails). Although the analysis of the data primarily indicated that participants who demonstrated higher sensation seeking were poor at discriminating between phishing and legitimate stimuli, the authors write that “Performance on the CRT also predicted susceptibility”.

A further motivation for our work was the tendency that many have of sharing of personal data when they do not have to. In the digital economy, we pay with our data [25], [26], [27]. For many applications, we have to give consent to sharing, but not always. All Internet-users can be targeted by phishing, and requests for sharing of data generally. We therefore chose to use national population samples rather than convenience samples or a sample with students only.

3. Research method

We carried out three surveys in Norway in cooperation with two different market research companies, to achieve our aim of national studies on an issue affecting a broad section of the population. The three surveys included questions from the Eurostat-survey about credit cards and misuse of data [28]. The formulation of these questions was discussed with the national bureau of statistics in Norway. This means that the findings in our studies, the demographical profile and the number that reported falling for phish can be compared to statistical data published by the national bureau of statistics.

The Cognitive Reflection Test was used to assess participants thinking styles, intuitive versus analytical. While in some countries many in the general public know with the correct answers to the CRT [29] the CRT has, as far as we know, not been used in a national, large-scale survey in Norway before. We also designed a behavioral measure concerning disclosure of personal data and demographics. We asked the participants for consent, to give us access to all the data about the participant that the market research company already had. Since the market research company was the data-processor, and we did not actually receive the data, we did not need ethical approval for the studies.

For the sample sizes we used the Eurostat-stat cybersecurity 2017 survey [28] as an indication. In this survey, 8 percent answered that they had experienced identity theft, that is someone stealing personal data and impersonating the person. On the basis of this we set a target of at least 100 respondents in each study who have experienced phishing.

The participants were recruited from two panels, citizens that are 18 years to 79 years in study 1 and 16 to 69 years in study 2 and 3. In total, study 1 had 1340 respondents 18 – 79 years old, and 1148 with the age 18 to 69. In study 2 there were in total 1405 individuals aged 16 to 69 years, and in study 3 1290 individuals. We excluded the 70 plus age group from study 1 in order to have a more similar age-profile for the three studies. The study 2 and the 3 participants were recruited from the same panel, but none of the study 2 participants were invited to participate in study 3.

3.1. Participants, the survey format and measurements

The participants received an email and answered the web-based survey on a PC or smart-phone, which took 10-15 minutes. For the CRT , with used the open format in study 1. In study 2, 50% received the open format, and 50% the multiple-choice format for the three CRT-items. In study 3, 75% received the open format, and 25% the multiple-choice format for the three CRT-items. The scores on the CRT [7] are reported in Table 3. For the three CRT-questions, the mean time used for the open format was 186 seconds vs. 108 seconds for the multiple-choice format.

In the first two studies 49% were male and 51% female, in the third study 50% male and 50% female.

Table 1. The age profile of the participants and descriptive statistics of the CRT.

Age:	16-19	20-29	30-39	40-49	50-59	60-69
Study1	2%	17%	22%	22%	21%	17%
Study2	8%	21%	16%	18%	20%	18%
Study3	9%	21%	18%	19%	18%	15%
	Min	Max	Mean	SD	Skewness	Kurtosis
Age Study1	18	69	43.9	14.1	0.01	-1.05
Age Study2	16	69	42.0	15.7	-0.02	-1.32
Age Study3	16	69	41.8	15.5	0.11	-1.18
CRT, Study1	0	3	1.25	1.76	0.30	-1.42
CRT; Study2	0	3	0.97	0.72	-0.01	-1.54
CRT; Study3	0	3	0.53	0.90	1.56	1.21

Table 1 shows that, persons of ages above 19 years were uniformly represented in our three samples. Skewness and Kurtosis are descriptive statistics for distribution. Skewness represents the extent to which scores have a tendency toward the upper or lower end of a distribution, while kurtosis indicates the extent to which a distribution of scores is relatively flat or relatively peaked. If the result is greater than +/- 2.0, the variable has a skewness problem. This is not the case for our studies, see Table 1.

In Table 2 we present the educational profile of the participants.

Table 2. Participants' educational profile. (Number of respondents in the parenthesis in the first column).

	Primary education	Secondary education	College & University, lower degree	University, higher degree
Study1 (1148)	7%	35%	38%	20%
Study2 (1405)	18%	36%	30%	17%
Study3 (1290)	13%	42%	29%	17%

The three measures used in the studies were the Cognitive Reflection Test, the three items developed by Frederick [7], a self-reported measure on phishing similar to the measurement used in the Eurostat-survey [28], and a behavioral measure on disclosure of personal data and demographics. The three CRT-items are often referred to as the bat/ball, widgets and the lilypad problem [13]. The open format, where the respondents fill in the answers themselves, is the standard CRT format. Recently a multiple-choice format has been developed. The motivation for using a multiple-choice format, using typical answers from studies with the open format, has been to save time for the respondents [30].

Table 3 presents the share of the respondents with all wrong answers, one correct, two correct and all three correct.

Table 3. The three CRT-items (Number of respondents in the parenthesis in the first column).

	All wrong	One correct	Two correct	All three correct
Study 1, open question	38%	21%	20%	22%
Study 2, open and multiple choice	43%	29%	16%	12%
Study 3, open and multiple choice	69%	15%	10%	6%

The context for our experiment on disclosure of personal data was that the participants in both studies had taken part in surveys before as panel members. The market research company has the answers to these surveys in their database but will normally not share this information with other clients. However, it is possible to link data and build a very detailed profile of each respondent based on answers to previous surveys. This was the context for our experiment on disclosure of personal data. We asked, in cooperation with the market research company, if we could have access to their answers to previous surveys and their Facebook profiles and with all these data build new profiles of them. The market research company, the data-processor, did not the share the personal data with us as client.

Both studies used two questions from the Eurostat-survey about credit cards and misuse of data. The formulation of these questions was discussed with the national bureau of statistics. In these questions, the word phishing is not used. Phishing is a term known in technological contexts, but its meaning is not known to all citizens. We performed a preliminary analysis of the understanding of the term. We used the question "can you, in your own words, describe was phishing is?" Of the respondents, 64% did not write anything, 23% wrote an explanation that we coded as correct, and 12% an explanation that we coded as incorrect. The two questions in the Eurostat-survey do not use the word phishing, but misuse and theft, see Table 4.

Self-reports on phishing and phishing incidents can be criticized. How accurate are these type of data? Greitzer et al. 2021 [43] did a large scale simulated phishing experiment with 6938 faculty and staff members of an American university. They write (p. 36): "In summary, our results suggest that, among numerous variables studied in this experiment, the best predictor of phishing susceptibility is having been phished before. Individuals who report having been successfully phished in the last 2 months are more likely to succumb to one or more of our phishing emails."

In the following we refer to phishing, those who have and those how have not fallen for phishes, based on the answers to these two questions.

Table 4. Credit-card misuse and ID-theft

	Has experienced misused of credit or debit card, the last 12 months	Has not experienced misused of credit or debit card, the last 12 months
Study1 (1148)	10%	90%
Study2 (1405)	14%	86%
Study3 (1290)	10%	90%
	Has experienced ID theft the last 12 months	Has not experienced ID theft the last 12 months
Study1 (1148)	7%	93%
Study2 (1405)	8%	92%
Study3 (1290)	8%	92%
	Has experienced misused of credit or debit card or ID-theft	Has experienced misused of credit or debit card or ID-theft
Study1 (1148)	12%	88%
Study2 (1405)	15%	85%
Study3 (1290)	14%	86%

Table 4 shows that around 10 percent of participants reported that they have experienced misuse, which is similar to the numbers reported in the Eurostat-surveys.

Table 5 presents the share of the respondents with all wrong answers, one correct, two correct and all three correct for the four educational groups.

Table 5. Education and the CRT

	All wrong	One correct	Two correct	All three correct
Study 1 Primary education (84)	56%	16%	16%	13%
Study 1 Secondary education (398)	48%	22%	17%	13%

Study 1 Univ. & college, lower level (432)	32%	23%	23%	22%
Study 1 Univ. & college, higher level (234)	24%	16%	20%	40%
Study 2 Primary education (246)	48%	31%	15%	7%
Study 2 Secondary education (498)	49%	28%	16%	8%
Study 2 Univ. & college, lower level (413)	40%	31%	15%	14%
Study 2 Univ. & college, higher level (244)	32%	27%	18%	23%
Study 3 Primary education (163)	77%	12%	8%	3%
Study 3 Secondary education (541)	71%	17%	9%	4%
Study 3 Univ. & college, lower level (372)	67%	13%	11%	9%
Study 3 Univ. & college, higher level (214)	64%	13%	14%	9%

3.1. Hypotheses – Sharing of Data and the CRT as a Predictor of Susceptibility

A low score on the cognitive reflection test indicates a tendency towards intuitive decision-making [8], [13]. Jones et al. [32], in their phishing study, found that performance on CRT predicted susceptibility to phishing. We hypothesized that education and CRT are predictors of falling for phishes as follows:

Hypothesis 1: Education is a predictor of susceptibility to phishing. In comparison to those with low education, those with high education are less susceptible.

There are many studies documenting that it is hard to detect phishing. Based on this we formulated the second hypothesis stating that an intuitive decision-making style measured with the CRT can predict falling for phishing.

Hypothesis 2: The CRT is a predictor of susceptibility to phishing. In comparison to those with a low score on the CRT, those with high score on the CRT are less susceptible to phishing.

Previous research has shown that females generally score lower on the CRT scores [7], [18] and so we expect them also to be more susceptible to phishing. However, studies on susceptibility to phishing did not find an effect of gender [32], [33]. Studies have indicated that in some situations, men take more risks than women [34].

Hypothesis 3: Gender is a predictor of susceptibility to phishing.

Hypothesis 4: Willingness to share personal data is a predictor of susceptibility to phishing.

4. Results

To test our hypotheses, we chose binary logistic regression with a dichotomous variable, 'has fallen for phish (yes/no)', as the dependent variable. One of the purposes was to investigate the question: is CRT score a good predictor of falling for phishing when we include the other factors gender, age, education and disclosure of data as variables?

Binary logistic regression is a form of regression used when the dependent variable is a dichotomy and the independent variables are of any type. It can be used to predict a categorical dependent variable on the basis of continuous and/or categorical independent variables, in our case whether or not someone reports that they has fallen for phishing in the past. By this method, the model is used for the prediction of the probability of the occurrence of the event by fitting data to a logistic curve. Cases with probabilities above a given numerical cut-off are accepted. We chose 0.12, 01.5 and 0.14 based on the percentages for falling for phish in the three datasets, see Table 4. The binary logistic, with the chosen cut-offs 1 is categorised as success whereas cases lower than this cut off value are classified as 0 (failure). This method is used to test the null hypothesis that a linear relationship does not exist between the predictor variables and the log odds of the criterion variable. Goodness-of-fit tests, such as the likelihood ratio test, are available as indicators of model appropriateness, as is the Wald statistic to test the significance of individual independent variables.

We tested our models with the SPSS-software, version 27. In logistic regression models, the Hosmer-Lemeshow test [35]. Archer et al. [36] is a goodness of fit test. Hosmer and Lemeshow recommend sample sizes greater than 400. A Hosmer-Lemeshow statistic of > 0.05 is often used to reject the null hypothesis that there is no difference, implying that the model's estimates fit the data.

Of our two models (Table 6) the first has p-value smaller than 0.05 and the second and third p-value larger than 0.05. The Nagelkerke *R2* is a pseudo R-square and it is impacted by how lopsided the split of dependent variables is. Even so it is often used to assess model adequacy [35]. The Nagelkerke *R2* was 13.3% for study 1, 19.6% for study 2 and 10.9% in study 3. Misuse of credit-card and ID theft were coded as one binary variable, see Table 4.

In the binary logistic model, we included gender, age, education, the CRT scores and the behavioral measure of data disclosure as variables. Table 5 shows that it was those with the longest education that performed best on the CRT-test. Since it has been shown that those with good mathematical skills or cognitive abilities often perform better on the CRT, we included an interaction effect of CRT and education in the model.

Table 6 - Overall fitting indices for the binary logistics regression model.

Model summary – study 1 (N=1148)			
-2 Log likelihood		Cox and Snell R square	Nagelkerke R square
Step 3	777.294	0.063	0.121
Hosmer and Lemeshow Test			
Chi-square		df.	Sig.
Step 3	19.875	8	0.011
Model summary – study 2 (N=1404)			
-2 Log likelihood		Cox and Snell R square	Nagelkerke R square
Step 3	1043.91	0.089	0.157
Hosmer and Lemeshow Test			
Chi-square		df.	Sig.
Step 3	9.566	8	0.297
Model summary – study 3 (N=1290)			
-2 Log likelihood		Cox and Snell R square	Nagelkerke R square
Step 3	955.648	0.060	0.109
Hosmer and Lemeshow Test			
Chi-square		df.	Sig.
Step 3	7.865	8	0.447

The *R squares* indicated that the variables in the equation contributed to predicting the dependent variable falling for phishing.

We used the Wald statistic to identify the significant variables in the model. The Wald statistic is the square of the t-statistic and gives equivalent results for a single parameter and can be used to test the significance of particular predictors in a statistical model. As the method for selecting how independent variables are entered into the analysis, we choice backward Wald. The method analyzes the predictor variables and picks the one that predicts the most on the dependent measure. In the backward method, all the predictor variables chosen are added into the model. Then, the variables that do not (significantly) predict anything on the dependent measure are removed from the model one by one. The backward method is generally the preferred method because the forward method might produce so-called suppressor effects. These suppressor effects occur when predictors are only significant when another predictor is held constant.

Table 7. Variables in the Equation

Variable code	Beta estimates	SE	**Wald**	df	Sig.	Exp (B)
Study 1 - Step 3						
Male=1 Female=0	-0.407	0,19	**4.49**	1	**0.03**	1.50
Age	-0.330	0.07	**21.97**	1	**0.00**	0.79
Data-Disclosure No=0, yes=1	-0.711	0.19	**13.46**	1	**0.00**	0.49
CRT	-0.383	0.09	**18.29**	1	**0.00**	0.68
Constant	0.270	0.32	0.73	1	0.39	1.31
Study 2 – Step 3						
Male=1 Female=0	0.385	0.17	**5.42**	1	**0.00**	1.47
Age	-0.21	0.01	**15.47**	1	**0.00**	0.98

Data-Disclosure No=0, yes=1	-1.406	0.18	**64.70**	1	**0.00**	0.25
CRT	-0.288	0.09	**10.99**	1	**0.00**	0.75
Constant	0.446	0.28	2,63	1	0.11	1.56
Study 3- step 3						
	Beta estimates	SE	**Wald**	df	Sig.	Exp (B)
Education	0.191	0.92	**4,272**	1	**0,04**	1.21
Age	-0.024	0.01	**16.803**	1	**0.00**	0.97
Data-Disclosure No=0, yes=1	-1.085	0.17	**41.267**	1	**0.00**	0.34
CRT	-0.283	0.12	**5.868**	1	**0.02**	0.75
Constant	-0.335	0.33	1.029	1	0.31	0.72

Education, and the interaction term education x CRT, are not variables in the equation for study 1 and 2 see Table 8. For study 3, gender is not a variable in the equation. In the final model in step 3 for study 1 and 2 (see Table 7) data-disclosure, CRT score and age all had high Wald estimates indicating that they made the biggest contribution to the model; that is, they are predictors for our outcome variable, falling for phishes. In study 3, data-disclosure and age are the two factors with high Wald estimates – see Table 7.

Table 8. Variables not in the equation

		Score	df	Sig.
Study 1 – Step 3				
	Education x CRT	0.323	1	0.570
	Education	0.203	1	0.652
	Overall statistics	0.327	2	0.849
Study 2 – Step 3				
	Education x CRT	0.047	1	0.829
	Education	0.011	1	0.916
	Overall statistics	0.417	2	0.812
Study 3 – Step 3				
	Male 1 Female 0	0.937	1	0.33
	Education x CRT	0.075	1	0.784
	Overall statistics	1.008	2	0.604

The Wald statistic estimates indicated that data disclosure behaviour, CRT scores and age were predictors of falling for phishing. In our model, see Table 7 (that included other variables), education was not a predictor of falling for phish in study 1 and 2, and a very weak predictor in study 3. Out conclusion is that hypothesis 1 was rejected. The second hypothesis was supported; in all three studies, CRT score was a predictor of falling for phish. The third hypothesis was partly supported; in the first two studies, the Wald estimates indicated a gender difference, with men being more susceptible to falling for phish than women. Hypothesis 4, willingness to share personal data, was supported. The respondents that gave consent seems more susceptibility to phishing than those that did not.

6. Discussion and concluding remarks

Our results confirmed the potential of using the CRT as a test for the likelihood of a person's susceptibility to phishing as reported by the citizens. The CRT provides a useful tool for identifying one of the characteristics of people who would benefit from advice or tuition to help avoid falling for these damaging confidence tricks. Willingness to share data was also associated with susceptibility to phishing.

Our findings indicate that the CRT can be used with samples drawn from a national population. CRT has been developed and validated with student samples and very few studies have used the CRT with ordinary citizens, as we did in the present studies. When a convenience sample is used, it may not be representative of the population at large so that the results are of limited generalizability. National studies might serve as a reference for other studies. This is also why we cooperated with the national bureau of statistics on the wording of the questionnaire.

However, it is much harder to design experiments with national samples, since the participants are not in a controlled environment. The time-factor plays a role in conjunction with the difficulty of the tasks. When a task takes many minutes, some participants will abandon it. Those with less education and other groups such as the elderly might behave differently from students. These can be recruited in a national sample. This is one of the reasons why the recommendation is that researchers should also use these samples. There are also ethical issues that are more challenging in uncontrolled environments, such as the issue of informed consent. There is also the issue of the expectations of the survey participants. They are used to answering questions, and less used to doing tasks and being tested in a study that includes the CRT problems. Some market research companies might hesitate to carry out studies that could attract complaints and negative publicity.

In the USA and some other English-speaking countries, the CRT is quite well known. If a respondent knows the correct answer in advance, the CRT cannot be used as intended. This is one of the reasons why alternatives to the standard CRT have been developed, tested and used in some recent studies [37]. In non-English speaking countries, such as the country of this study, Norway, it has rarely been used, so that it is unlikely that respondents will know the answers already. Others have studied the repeated exposure effect of the CRT and concluded that the CRT is robust to multiple testing and is stable across time [42]. However, if someone performs an online search, he or she will be able to find the correct answers easily.

The present results demonstrate that those with more education perform significantly better than others on the CRT. One of the strengths of using the CRT is that it is not a self-reported measurement but rather, assuming that the respondent does not search for the answer (or know the answer in advance), tells us about the respondent's individual behavior and characteristics. Our study indicates that individual citizens can perform well on the CRT without higher education. In our logistics models that included demographics, a measure on data-sharing and the CRT, it was the data-sharing behavior and the CRT that contributed significantly to predicting susceptibility to phishing, not demographics.

Sirota & Juanchich [38] argue that the standard open format should be replaced by a multiple-choice format because it is less likely that someone will perform a search to find the answer; instead the respondent will give a spontaneous response. However, the comparison we did of the two formats in the study indicated that, including for the multiple-choice format, some users took a very long time to answer the three questions. For the three CRT questions the mean completion time was 186 seconds for the open format vs 108 for the multiple-choice format. A solution would be to use a timer. After x seconds, the next section or question is presented, and in the case of no answer being given this will be recorded as a no answer or a wrong answer. This approach was used by Da Silva et al. [39] and should be considered for future studies with the CRT.

It is important to mention that we do not know that the respondents reported honestly when they answered the questions and that we actually know whether they have actually fallen for phishing or not. We speculate that some that spent long times on the CRT questions had searched for the answers online. In a controlled laboratory setting it is less likely that this will happen. When a respondent is answering a survey on his or her PC or smartphone, he or she may be distracted and may not really care much about the questions and the answers given [40]. In a lab., a class-room or other controlled environment this is less of a problem. Another drawback of large-scale surveys is that it is costly to recruit many respondents and run a large-scale study in a population.

The CRT is useful for research on why online users fall for phishing but is not the only measure that can be recommended. We opted for a measure on data disclosure in our studies to complement CRT, as well as demographics. For future research, we suggest that the CRT should be used in actual or semi-natural phishing experiments, together with other measurements of risk propensity [41] inattention, optimism bias or overconfidence.

Acknowledgements

This research was supported by Research Council Norway, the grants 270969 and 310105 (NORCICS). The author gratefully acknowledges John Waterworth for proofreading the draft of the manuscript.

10. References

[1] Jones, Jagatic, T., Johnson, N., Jakobsson, M., and Menczer, F. Social phishing. *Communications of the ACM*, 2007. 5(10), 94–100.

[2] Dhamija, R., Tygar, J. D., and Hearst, M. Why phishing works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada, April 22 - 27, 2006. R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, and G. Olson, Eds. CHI '06. ACM Press, New York, NY, 581-590.

[3] Acquisti, A. Privacy in electronic commerce and the economics of immediate gratification. In: EC '04 Proceedings of the 5th ACM Conference on Electronic Commerce, USA, 2004. 21-29.

[4] Barnes, S. B. A privacy paradox: Social networking in the United States. First Monday, 2006. 11(9). Retrieved from http://firstmonday.org/article/view/1394/1312.

[5] Acquisti, A., Adjerid, R. Balebako, L., Brandimarte, L. Cranor, S., Komanduri, P. Leon, N. Sadeh, F. Schaub, M. Sleeper, Y. Wang, and S. Wilson. Nudges for privacy and security: Understanding and assisting users choices online. ACM Computing Surveys, 2017. Vol. 50 (3) Article 44, Aug.

[6] Nicholson, J., Coventry, L., and Briggs, P. 2017. Can we fight social engineering attacks by social means? Assessing social salience as a means to improve phishing detection. In Proceedings of the Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017). Santa Clara, CA: USENIX.

[7] Frederick, S. Cognitive Reflection and Decision Making. Journal of Economic Perspectives 2005. 19(4): 25-42.

[8] Toplak, M. E., West, R. F., and Stanovich, K. E. The Cognitive Reflection Test as a predictor of performance on heuristics and biases tasks. *Memory & Cognition*, 2011. 39, 1275–1289.

[9] Ferreira, A. and Vieira-Marques, P. Phishing through time: A ten year story based on abstracts, Proceedings of the 4th International Conference on Information Systems Security and Privacy, 2018. vol. 1, pp. 225-232.

[10] Volkamer. M, Renaud, K., Reinheimer, B. and A. Kunz, "User experiences of TORPEDO: TOoltip-powered phishing email DetectiOn ", *Computers & Security*, February 2017.

[11] Stockhardt, ., Reinheimer, B., Volkamer, M., Mayer, P., Kunz, A., Rack, P., and Lehmann, D. Teaching phishing-security: Which way is best? 2016. *Vol. 471. 31st IFIP TC 11 International Conference on Systems Security and Privacy Protection, SEC 2016* (pp. 135-149): Springer New Y. LLC.

[12] Kahneman, D. Thinking, fast and slow. 2011. New York, NY: Farrar, Straus and Giroux.

[13] Toplak, M. V., West, R. F., and Stanovich, K. E. Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning,* 2014. 20, 147–168.

[14] Pennycook, G., Cheyne, J. A., Koehler, D. J., and Fugelsang, J. A. Is the cognitive reflection test a measure of both reflection and intuition? *Behav. Res. Methods*, March 2016, Volume 48, Issue 1, pp 341–348.

[15] Pennycook, G, R. and D. Lazy, Not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, June 20, 2018, 10.1016/j.cognition.2018.06.011.

[16] Bialek, M., and Pennycook, G. The Cognitive Reflection Test is robust to multiple exposures. Behavior Research Methods. 2018.

[17] Mata, A., Ferreira, M. B., and Sherman, S. J. The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology*, 2013, 105, 353–373.

[18] Campitelli, G., and Gerrans, P. Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory and Cognition*, 2014. 42 (3), 434–447.

[19 Cokely, E. T., and Kelley, C. M. Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making*, 2009. 4, 20–33.

[20] Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., and Pardo, S. T. Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making,* 2012. 25, 361–381.

[21] Kumaraguru, P., Rhee, Y., Sheng, S. et al. Getting users to pay attention to anti-phishing education: Evaluation of retention and transfer. In Proceedings of the Anti-Phishing Working Group's Second Annual eCrime Researchers. 2017.

[22] Butavicius, M., Parsons, K. Pattinson, M. and A. McCormac, Breaching the Human Firewall: Social engineering in Phishing and Spear-Phishing Emails, ArXiv160600887, May 2016.

[23] Petraityite, M., Dehghantanha, A., and Epiphaniou, G. 2017. Chapter 6 - Mobile Phone Forensics: An investigative framework based on user impulsivity and secure collaboration errors. In Contemporary Digital Forensic Investigations of Cloud and Mobile Applications (79– 89). Syngress.

[24] Jones, H. S, Towse J. N, Race N, and Harrison T. Email fraud: The search for psychological predictors of susceptibility. 2019. PLoS ONE 14(1): e0209684.

[25] Elvy, S. A. Paying for Privacy and the Personal Data Economy. Columbia Law Review, 2017, 117 (6): 1369–459.

[26] Hacker, P. and Petkova, B. Reining in the big promise of big data: transparency, inequality, and new regulatory frontiers, *Northwestern Journal of Technology and Intellectual Property*, 2017, 15:1-42.

[27] Greengard, S. Weighing the impact of GDPR. *Communications of the ACM*, 2018, 61(11): 16–18.

[28] European Union, 1. 2017. 5661. Special Eurobarometer 464a "European attitudes towards cyber security." September 2017.

[29] McCall, R. Can you pass the world's shortest IQ test? It's just three questions long, but few can get them all right. 2017, Accessed April 14. 2019 from: http://www.iflscience.com.

[30] Šrol, J. Dissecting the expanded cognitive reflection test: an item response theory analysis. *Journal of Cognitive Psychology*, 2018, 30:7, 643-655.

[31] Jones, H. S, Towse J. N, Race N, and Harrison T. Email fraud: The search for psychological predictors of susceptibility, 2019. PLoS ONE 14(1): e0209684.

[32] Jones, H. What makes people click: assessing individual differences in susceptibility to email fraud. 2016, eprints.lancs.ac.uk.

[33] Parsons, K., McCormac, A. Pattinson, M., Butavicius, M., and Jerram, C. Phishing for the truth: A scenario-based study of users' behavioural response to emails. In IFIP International Information Security Conference (pp. 366-378). 2013. Berlin Springer

[34] Charness, G., and Gneezy, U. Strong evidence for gender differences in risk-taking. *Journal of Economic Behavior and Organization*, 2012, 83, 50–58.

[35] Hosmer, W. and Lemeshow, S. Applied logistic regression. 1989. New York: Wiley.

[36] Archer, K. J., Lemeshow, S., and Hosmer, D. W. (2007). Goodness of fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics & Data Analysis*, 51, 4450–4464

[37] Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., and Hamilton, J. The development and testing of a new version of the cognitive reflection test applying item response theory. *Journal of Behavioral Decision Making*, 2016, 29. 453–469.

[38] Sirota, M., and Juanchich, M. Effect of response format on cognitive reflection: Validating a two- and four-option multiple choice question version of the Cognitive Reflection Test. *Behavior Research Methods*. 2018. doi: 10.3758/s13428-018-1029-4.

[39] DaSilva, S, Da Costa Jr N, Matsushita R, Vieira C, Correa A, and De Faveri D. Debt of high-income consumers may reflect leverage rather than poor cognitive reflection, *Review of Behavioral Finance*. 2017.

[40] MacKenzie, S. B. and P. M. Podsakoff. Common Method Bias in Marketing: Causes, Mechanisms, and Procedural Remedies. *Journal of Retailing* 2012. 88: 542-555.

[41] Lejeuz, C. W., Jennifer P. Read, Christopher W. Kahler, Jerry B. Richards, Susan E. Ramsey, Gregory L. Stuart, David R. Strong, and Richard A. Brown. Evaluation of a Behavioral Measure of Risk Taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied* 2002, 8(2):75-84.

[42] Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018) Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision Making*, 13, 260–267.

[43] Greitzer, F. L., Li, W., Laskey, K. B., Lee J., and Purl, J. 2021. Experimental Investigation of Technical and Human Factors Related to Phishing Susceptibility. ACM Trans. Social Computing. 4, 2, Article 8 (June 2021)