Exploring Susceptibility to Phishing: the Cognitive Reflection Test and Other Possible Predictors

Ingvar Tjostheim
Norwegian Computing Center, P.O. Box 114 Blindern,
N.O.-0314 Oslo, Norway
ingvar@nr.no

John A. Waterworth
Umeå University, Main Campus,
SE-901 87 Umeå, Sweden
jwworth@informatik.umu.se

Abstract

The research objective of this study was to investigate factors contributing to phishing susceptibility, expanding on findings from previous studies. We report results based on five, large-scale surveys of national populations from which we collected data about cognitive strategies using the Cognitive Reflection Test (CRT), privacy attitudes, data disclosure behaviors, and demographic variables. We used binary logistic regression to analyze the relationship between these factors and susceptibility to phishing attacks. We found that willingness to share personal data and CRT scores significantly predicted phishing susceptibility. Younger people were somewhat more susceptible than older age-groups. as were males than females. Importantly, these findings suggest that phishing susceptibility is not simply a function of cognitive ability, but also of individual differences in privacy attitudes and data disclosure behaviors. Their credibility is enhanced by the use of five large-scale studies with national populations, unlike earlier studies primarily relying on smaller-scale student populations.

1. Introduction

Phishing refers to a deceptive tactic employed to trick unsuspecting individuals into divulging their personal information online, allowing the perpetrator to fraudulently exploit their credentials (Jones et al. 2007, Dhamija et al. 2006). Phishing attacks are often highly sophisticated, which means that even well-educated and cautious internet users can be vulnerable to falling victim to such attacks. There exists a tendency among many internet users to engage in behavior that compromises their privacy, indicating a disparity between their privacy attitudes and actual conduct (Barnes, 2006, Acquisti, 2004). Nicholson et al. (2017) argue that phishing is an example where users demonstrate overconfidence, and other contributing factors include inattention, optimism biases, irrational behavior, limited cognitive resources, as well as various biases and heuristics that are well-recognized by behavioral researchers (Acquisti, 2004:32). Our study targeted a national population by recruiting participants through two professional market research firms. The study encompassed inquiries regarding phishing and the misuse of personal data, along with a choice experiment concerning the sharing of personal information. To gauge individuals' ability to suppress intuitive and impulsive incorrect responses in favor of reflective and accurate answers, we employed the Cognitive Reflection Test. According to Toplak et al. (2011) the CRT possesses the potential to serve as a distinct predictor of performance on various tasks involving heuristics and biases.

2. Related work and motivation for the research

Ferreira & Vieira-Marques (2018) provide an overview of phishing research spanning a decade, based on an analysis of 605 scientific journal abstracts. They conclude that there is no singular solution to mitigate the phishing threat and advocate for future research to focus on socio-technical and integrated approaches that encompass a comprehensive understanding of both human-computer interaction and user-specific characteristics. Assessing these unique user characteristics served as a primary motivation for our research.

Volkamer et al. (2017) and the APWG Internet Policy Committee Global Phishing Survey reveal that, on average, it takes 28.75 hours to detect new phishing websites. Users remain largely vulnerable to phishing attacks until malicious websites are identified and blocked Stockhardt (2016). In order to avoid falling victim to phishing during this period, users must engage in reflective decision-making rather than mere compliance with requests. This serves as the impetus behind our investigation into the interplay between intuitive (automatic decision-making) and reflective problem-solving styles concerning susceptibility to phishing and willingness to share personal information,

URI: https://hdl.handle.net/10125/106955
978-0-9981331-7-1

which is why we incorporated a version of the CRT into our studies.

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

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). 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 (Toplak et al. 2011, Toplak et al. 2014). To score highly on the CRT, the respondent needs to reflect on and question their initial intuitive responses (Pennycook, 2016, Pennycook & Lazy, 2018) 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 (2018) 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)."

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 (Campitelli & Gerrans, 2014, Cokely & Kelly, 2009, Liberali et al. 2012, Toplak et al. 2011, Toplak et al., 2014).

The CRT aims to cue intuitions that are common across people and lead to the same 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. (2017) 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. (2016) performed a phishing study with 121 students. These researchers found a significant negative correlation between CRT scores and link safety judgments for spear-phishing (ρ = -.23, p = .014, N = 112) and phishing (ρ = -.3, p = .001, N = 112) emails, but no significant correlation between performance on the CRT and link safety judgments on genuine emails (ρ = -.01, p = .973, N = 114). Petraityte et al. (2017) 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. (2019) 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 personal data when they do not have to. In the digital economy, we often pay with our data (Elvy, 2017, Hacker & Petkova, 2017, Greengard, 2018). 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 commissioned five surveys in cooperation with three different market research companies, to achieve our aim of national studies on an issue affecting a broad section of the population. The five surveys included questions from the Eurostat-survey about credit cards and misuse of data (European Union, 2017). 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. After two questions about whether the respondents had experienced credit-card misuse and, or ID-theft, they were asked to report a phishing incident by writing a sentence about what they had experienced. Giving an example was optional. The majority skipped the question, but many types of phishing incidents were

reported. The Cognitive Reflection Test was used to assess participants' thinking styles, intuitive versus analytical. While in some countries many in the general public know the correct answers to the CRT (McCall, 2021) the CRT has, as far as we know, not been used in a similar national 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, consent for receiving and analyzing personal data.

For the sample sizes we used the Eurostat-stat cybersecurity 2017 survey (European Union, 2017) 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 by the market research companies IPSOS AS, Poling & Statistics AS, and Norstat AS. Citizens from 18 years to 79 years old, participated in Study 1. In the next four studies citizens 16 to 69 years were recruited. The numbers of respondents were; 1148 in study 1, 1405 in study 2, 1290 in study 3, 1630 in study 4 and 1882 in study 5. The data were collected from November 2019 to March 2023. The data from the first three studies was used in a similar examination and published in a HICSS-paper by Tjostheim (2022).

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. For the CRT, we 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. In study 4 and 55% received the open format, and 50% the multiple-choice format. The scores on the CRT [7] are reported in Table 3a and Table 3b. In study 2 we measured the time the respondents spent on the CRT. 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 studies 1, 2 and 4 49% were male and 51% female, in the third study 50% were male and 50% female, in the fifth study 46% were male and 54% female.

Table 1. The age profile of the participants. (Number of respondents in the parenthesis in the first column).

Age:	18-19	20-29	30-39	40-49	50-59	60-69
Study 1 N=1148	2%	17%	22%	22%	21%	17%
	16-19	20-29	30-39	40-49	50-59	60-69
Study 2 N=1405	8%	21%	16%	18%	20%	18%
Study 3 N=1290	9%	21%	18%	19%	18%	15%
Study 4 N=1630	7%	8%	23%	23%	14%	18%
Study 5 N=1882	11%	10%	24%	20%	20%	15%

Table 1 shows that persons of ages above 19 years were uniformly represented in four of the five samples. Study 4 had fewer respondents under 30 years old. In Table 2 we present the educational profile of the participants.

Table 2. Participants' educational profile

	Primary education	Secondary education	College & University, lower	University, higher degree
Study 1	7%	35%	38%	20%
Study 2	18%	36%	30%	17%
Study 3	13%	42%	29%	17%
Study 4	9%	42%	25%	24%
Study 5	12%	38%	25%	25%

The three measures used in the studies were the Cognitive Reflection Test, with the three items developed by Frederick (2005), a self-reported measure on phishing similar to the measurement used in the Eurostat-survey (European Union, 2017), and a behavioral measure on disclosure of personal data and demographics. 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 Šrol (2018).

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

	All wrong	One correct	Two correct	All three correct
Study 1, open q. Study 2, open and	38%	21%	20%	22%

multiple choice	43%	29%	16%	12%
Study 3, open and multiple choice	69%	15%	10%	6%
Study 4, open and multiple choice	26%	29%	22%	27%
Study 5, open and multiple choice	27%	45%	26%	3%

In both study 3 and 5 only a small fraction, 6% and 3% respectively, provided the correct responses to all three questions. The reason for this outcome is not entirely clear, except for the fact that it is often effortless to provide an incorrect answer to a CRT-question. To answer such questions accurately, one needs to be attentive, pause, and reflect before responding.

The original three CRT-question by Frederick (2005:27) can be referred to as the classic three numerical CRTs.

(1) A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? _____ cents

(2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____ minutes

(3) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? _____ days

Figure 1 The three CRT-questions, Frederick (2005)

The first question is named the ball and the bat. In Norwegian surveys 1 to 4 a tennis racket and a tennis ball were used since baseball is not a common sport in Norway. In the fifth survey, a burger and French fries was used. The next two, the widget question and the Lilypad question were translated into Norwegian.

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 given in these previous surveys in their database but will not share this information with other clients. The contract between the market research company and the members of the panel explicitly stated that answers to one survey for one client would not be not shared with other clients of the market research company. The market research company has a demographic profile of each panel-member, but the answers to each survey are not linked to this profile. 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 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 about credit card or debit card misuse, and ID-theft.

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
Study 1	10%	90%
Study 2	14%	86%
study 3	10%	90%
Study 4	8%	92%
Study 5	6%	94%
	Has experienced ID theft, the last 12 months	Has not experienced ID theft, the last 12 months
Study 1	7%	93%
Study 2	8%	92%
Study 3	11%	92%
Study 4	11%	89%
Study 5	12%	88%
	Has experienced misused of credit or debit card or ID-theft	Has experienced misused of credit or debit card or ID-theft
Study 1	12%	88%
Study 2	15%	85%
Study 3	14%	86%
Study 4	13%	87%
Study 5	16%	84%
Average for the 5 studies	15%	85%

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

3.2. 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 (Toplak et al. 2011, Toplak et al., 2014, Jones et al. 2019). Jones et al. (2019), 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: Individuals with higher levels of education are less susceptible to phishing attempts compared to those with lower levels of education.

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 (Frederick, 2005. Campitelli & Gerrans, 2014) 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 (Parsons et al, 2013, Jones, 2016). Studies have indicated that in some situations, men take more risks than women (Charness & Gneezy, 2012). Our third hypothesis concerns the possibility of gender differences in susceptibility to phishing, as follows:

Hypothesis 3: Men are more susceptible to phishing than females.

In their responses to the market research company, our participants were asked to provide access to their answers to previous surveys and their Facebook profiles. Our fourth hypothesis was based on the *a priori* assumption that people who were willing to share their personal data are more likely to fall for phishing, as follows:

Hypothesis 4: Those more willing to share personal data are more susceptible to phishing than those less willing to share personal data.

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 our objectives was to explore the following question: Does the CRT score serve as a reliable indicator for predicting susceptibility to phishing incidents when considering other variables such as gender, age, education, and data disclosure? Alternatively, is the willingness to share data, specifically data-disclosure, an equally strong or superior predictor?

The Kruskal-Wallis test in 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 (Sinayev & Peters, 2015), we included an interaction effect of CRT and education in the model. For the calculation of the effect size correlations, we refer to Tabachnick & Fidell (2013) and Menard, (2000) the video by Crowson in 2021.

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 was not the case for our studies. For age, skewness varied from -0.15 to 0.10 and kurtosis from -1.32 to -0.93. For the CRT, skewness varied from -0.01 to 1.56 and kurtosis from -1.54 to 1.21.

Table 5. Education and the CRT

	All wrong	One correct	Two correct	All three correct
Study 1 Primary e.	56%	16%	16%	13%
Study 2	48%	31%	15%	7%
Study 3	77%	12%	8%	3%
Study 4	32%	23%	25%	21%
Study 5	28%	40%	28%	4%
Study 1	48%	22%	17%	13%
Secondary	49%	28%	16%	8%
Study 2	71%	17%	9%	4%
Study 3	29%	29%	20%	22%
Study 4	32%	49%	18%	2%
Study 5				
Study 1 University & college, lower	32%	23%	23%	22%
Study 2	40%	31%	15%	14%
Study 3	67%	13%	11%	9%
Study 4	23%	29%	23%	26%
Study 5	24%	41%	32%	3%
Study 1 University & college, higher	24%	16%	20%	40%
Study 2	32%	27%	18%	23%
Study 3	64%	13%	14%	9%
Study 4				
Study 5				

	23% 21%	17% 43%	11% 34%	38% 3%
Kruskal-Wallis H-test (one-way Anova)				
	Kruskal-Wallis H		df	Assym p.Sig
Study 1	78.431		3	<.001
Study 2	44.799		3	<.001
Study 3	18.006		3	<.001
Study 4	38.582		3	<.001
Study 5	46.020		3	<.001

Binary logistic regression is a form of regression used when the dependent variable is a dichotomy and the independent variables are of any type. Binary logistic regression 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 s/he 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, 0.15, 0.14, 0.13 and 0.16 based on the percentages for falling for phish in the datasets. 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 29. In logistic regression models, the Hosmer-Lemeshow test (Hosmer & Lemeshow, 1989) is used. Archer et al. (2007) 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. For four of the five studies, this criterion was met - see tables 6 and 7.

In the binary logistic model, we included gender, age, education, the CRT scores and the behavioral measure of data disclosure as variables. Misuse of credit-card and ID theft were coded as one binary variable.

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

Model summary			
-2 Log likelihood		Cox and Snell R square	Nagelkerke R square
Study 1, step 3	771.073	0.065	0.124
Study 2, step 3	1045.87	0.090	0.158
Study 3, step 3	959.454	0.057	0.105
Study 4, step 3	1131.92	0.012	0.023
Study 5, step 2	1594.77	0.014	0.024
Hosmer and Lemeshow Test			
Chi-square		df.	Sig.
Study 1, step 3	22.062	8	0.005
Study 2, step 3	6.207	8	0.624
Study 3, step 3	12.282	8	0.139
Study 4, step 3	9.875	8	0.329
Study 5, step 2	13.065	8	0.110
Effect size correlations for assessing global fit			
Study 1	0.31		
Study 2	0.34		
Study 3	0.27		
Study 4	0.11		
Study 5	0.12		
The five studies combined	0.21		

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.

In the final model in step 3 or 2 for study 1, 2, 3 and 5 (see Table 7) data-disclosure had high Wald estimates. For study 1 and 2 the CRT had high Wald estimates. For study 1, 2 and 3 age had high Wald estimates.

Table 7. Variables in the Equation, step 3 (St1=study 1, St2 =study 2, St3 =study 3 etc.)

Variable code	Beta est.	SE	**Wald**	**Sig.**	Exp (B)
St1 Female=0					
Male=1	0.39	0.19	**3.95**	**0.04**	1.46
St2 - " -	0.38	0.17	**5.17**	**0.02**	1.46
St3 - " -	-	-	**-**	**-**	-
St4 - " -	0.52	0.15	**11.59**	**0.00**	1.69
St5 - " -	0.24	0.13	3.47	0.06	1.36
St.1 Age	-0.33	0.07	**22.12**	**0.00**	0.97
St.2 - " -	-0.21	0.01	**15.30**	**0.00**	0.98
St.3 - " -	-0.02	0.01	**16.037**	**0.00**	0.98
St.4 - " -	-	-	-	-	-
St.5 - " -	-0.01	0.00	**5.47**	**0.02**	0.99
No=0, Yes=1					
St1 Data-Disclosure	0.73	0.20	**13.91**	**0.00**	2.07
St2 - " -	1.40	0.18	**64.11**	**0.00**	4.05
St3 - " -	1.10	0.17	**41.86**	**0.00**	3.00
St4 - " -	0.34	0.17	**4.16**	**0.04**	1.40
St5 - " -	0.52	0.13	**15.09**	**0.00**	1.69
St1 CRT	-0.39	0.09	**19.11**	**0.00**	0.68
St2 - " -	-0.32	0.09	**13.16**	**0.00**	0.73
St3 - " -	-0.18	0.10	2.92	0.09	0.84
St4 - " -	-0.13	0.07	**4.01**	**0.04**	0.88
St5 - " -	-0.32	0.17	3.67	0.06	0.73
St3 Education	0.18	0.94	3.62	0.06	1.20
St4 - " -	0.11	0.07	3.00	0.08	1.12
St5 CRT x Education	0.10	0.05	**4.04**	**0.04**	1.11
St1 Constant	-0.66	0.32	**4.31**	**0.04**	0.56
St2 - " -	-1.22	0.24	**24.69**	**0.00**	0.30
St3 - " -	-1.70	0.32	**28.86**	**0.00**	0.18
St4 - " -	-2.36	0.22	**120.50**	**0.00**	0.09
St5 - " -	-0.90	0.32	**7.78**	**0.01**	0.41

Education is not a significant predictor – see Table 7. However, in study 5 the interaction term education with CRT is one of the variables in final model. for study 5 only – see Table 7. For study 3 and 4, gender is not a variable in the equation.

Table 8 - Overall fitting indices for the binary logistics regression combined model

Model summary – 5 studies combined 1 (N=7355)			
-2 Log likelihood		Cox and Snell R square	Nagelkerke R square
Step 2	5683.81	0.035	0.063

Hosmer and Lemeshow Test			
	Chi-square	df.	Sig.
Step 2	16.412	8	0.037
	Effect size correlations for assessing global fit		
The five studies combined	0.21		

The Wald statistic estimates indicated that data disclosure behaviour and CRT scores were predictors of falling for phishing – they are in all 5 models. In three of the models age is also a predictor - the younger fall for fish more often than the elderly. Table 7 shows that education was not a significant predictor of falling for phish in any of the five studies, but a weak predictor in study 3 and 4. In study 5, education x CRT was a significant predictor.

Our conclusion is that Hypothesis 1 was rejected. Hypothesis 2 was supported; in all five studies, CRT score was a predictor of falling for phish. Hypothesis 3 was partly supported; in the four of the 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. It was only in study 4 that it was not a strong predictor.

The questions and the demographical variables were unchanged from 2019 to 2023. We therefore merged the dataset and performed an analysis on a large dataset including all 7355 respondents. The Wald estimates in this case indicated that data-disclosure and CRT were strong predictors (see Table 9).

Table 9. Variables in the Equation

Variable code	Beta est.	SE	**Wald**	Sig.	Exp (B)
Five Studies combined - Step 2					
Male=1 Female=0	0.358	0.07	**26.055**	**0.00**	1.43
Age	-0.017	0.00	**50.704**	**0.00**	0.98
Data-Disclosure (no=0, yes1)	0.762	0.07	**112.39**	**0.00**	2.14
Education	0.107	0.04	**9.028**	**0.00**	1.11
CRT	-0.224	0.03	**50.539**	**0.00**	0.80
Constant	-1.599	0.13	**150.38**	0.00	0.20

This finding corresponds to the conclusions of the four hypotheses given above. The result emerges even more clearly when the analysis is done with all five studies together.

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. The CRT provides a useful tool for identifying some of the 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, and it played an even more significant role than CRT. However, for data disclosure a test such as the CRT has not been developed and validated.

Our findings indicate that both willingness to share data and 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. This is one of the reasons why the recommendation is that researchers should also use these types of samples.

There are also ethical issues that are more challenging in uncontrolled environments, such as the issue of informed consent. Then there is the issue of what the survey participants expect or agree to do. They are used to answering questions, and less used to doing tasks and being tested in a study that includes the CRT. 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 (Tabachnick & Fidell, 2103). In non-English speaking countries, such as the country of this study, Norway, it has rarely been used outside universities. 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 the highest education perform slightly better than the lowest educational group (primary education) on the CRT, but the highest education group is in a population much smaller than other groups. 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 (2018) 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. We received log-data from the survey-software to calculate the time used for the three questions. 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 so that, 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. (2017) 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 do not actually know whether they have actually fallen for phishing or not, but many of them wrote in a text box about an incident that had happened. 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 (MacKenzie & Podsakoff, 2012). In a lab., a class-room or other controlled environment this is less of a problem. Another drawback of large-scale surveys with professional market research companies is the costs of recruit respondents. Often researchers do not have a budget for this type of data-collection.

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 (Lejeuz et al. 2002) inattention, optimism bias or overconfidence.

Acknowledgements

This research was supported by Research Council Norway under the grant 310105 (NORCICS).

10. References

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

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

Archer, K. J., Lemeshow, S., & 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.

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

Bialek, M., & Pennycook, G. (2018). The Cognitive Reflection Test is robust to multiple exposures. *Behavior Research Methods*.

Butavicius, M., Parsons, K., Pattinson, M., & McCormac, A. (2016). Breaching the Human Firewall: Social engineering in Phishing and Spear-Phishing Emails, ArXiv160600887.

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

Charness, G., & Gneezy, U. (2012). Strong evidence for gender differences in risk-taking. *Journal of Economic Behavior and Organization*, 83, 50-58.

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

Crowson, M. (2021). A super-easy effect size for evaluating the fit of a binary logistic regression using SPSS - YouTube.com/watch?v=LA5gWgfpHkY.

DaSilva, S., Da Costa Jr., N., Matsushita, R., Vieira, C., Correa Am, & De Faeri, D. (2017). Debt for high-income consumers may reflect leverage rather than poor cognitive reflection. *Review of Behavioral Finance*.

Dhamija, R., Tygar, J. D., & Hearst, M. (2006). Why phishing works. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06) (pp. 581-590). ACM Press.

Elvy, S. A. (2017). Paying for Privacy and the Personal Data Economy. *Columbia Law Review*, 117(6), 1369-1459.

European Union. (2017: 5661). Special Eurobarometer 464a "European attitudes towards cyber security." September 2017.

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

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

Greengard, S. (2018). Weighing the impact of GDPR. *Communications of the ACM*, 61(11), 16-18.

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

Hosmer, W., & Lemeshow, S. (1989). Applied logistic regression. New York: Wiley.

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

Jagatic, T., Johnson, N., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, 50(10), 94-100.

Jones, H. (2016). What makes people click: Assessing individual differences in susceptibility to email fraud, eprints.lancs.ac.uk.

Kahneman, D. (2011). Thinking, Fast and Slow. New York, NY: Farrar, Straus and Giroux.

Kumaraguru, P., Rhee, Y., Sheng, S., et al. (2017). 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.

Lejeuz, C. W., Jennifer P. Read, Christopher W. Kahler, Jerry B. Richards, Susan E. Ramsey, Gregory L. Stuart, David R. Strong, & Richard A. Brown (2002). Evaluation of a Behavioral Measure of Risk Taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, 8(2), 75-88.

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

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

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

McCall, R. (2017). Can you pass the world's shortest IQ test? It's just three questions long, but few can get them all right. Accessed June 1, 2021, from https://www.rd.com/article/worlds-shortest-iq-test/.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54, 17-24.

Nicholson, J., Coventry, L., & 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.

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

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fuglesang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods*, 48(1), 341-348.

Pennycook, G., & Rand, D. (2018). Not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*.

Petraityte, M., Dehghantanha, A., & 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 (pp. 79-89).

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

Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, 6, 532.

Sirota, M., & Juanchich, M. (2018). 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*.

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

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

Tabachnick, B.G., & Fidell, L.S. (2013). Using multivariate statistics (6th ed). Pearson: Upper Saddle River, NJ.

Tjostheim, I. (2022). Phishing, Data-Disclosure and The Cognitive Reflection. Proceedings of the 55th Hawaii International Conference on System Sciences, URI: https://hdl.handle.net/10125/80268

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

Toplak, M. V., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20, 147-168.*

Volkamer, M., Renaud, K., Reinheimer, B., & Kunz, A. (2017). User experiences of TORPEDO: TOoltip-powered phishing email DetectiOn. Computers & Security.