An Examination of the Calibration and Resolution Skills in Phishing Email Detection

Full Papers

Yuan Li
Columbia College, Columbia, SC, USA
yli@columbiasc.edu

Jingguo Wang
University of Texas, Arlington, TX, USA
jwang@uta.edu

H. Raghav Rao
State University of New York, Buffalo, NY, USA
mgmtrao@buffalo.edu

Abstract

This study examines individuals' calibration and resolution skills in phishing email detection and tests the effects of several factors on both skills. It shows that calibration and resolution are two distinct capabilities of a person to detect phishing emails, and they are subject to the impacts of different factors: while calibration is influenced mostly by task factors such as familiarity with the emails, time to judgment, variability of time to judgment, and task easiness, resolution is influenced by both task factors such as variability of time to judgment and familiarity with the entity in the email, and individual characteristics such as online transaction experience and prior victimization of phishing attacks. The theoretical implication of the study is addressed, and the practical implication for designing effective training programs to improve one's phishing detection ability is also discussed.

Keywords

Phishing email detection, calibration skills, resolution skills, judgmental confidence, phishing detection accuracy

Introduction

In the age of cybersecurity, human is an integral part and the last line of defense against phishing attacks, as current legislative and technological solutions do not sufficiently prevent phishing emails from reaching one's inbox. Correspondingly, the capability of a person to correctly identify phishing emails from genuine business emails is a primary focus of research in the phishing detection literature. A number of studies have examined individuals' phishing detection ability and explored ways to improve detection accuracy (Vishwanath et al 2011; Wright and Marett 2010). Education programs such as "PhishGuru" and "Anti-Phishing Phil" were also created to teach users how to recognize cues to avoid falling for phishing attacks (Kumaraguru et al 2010).

While improving one's detection accuracy is important, it may not be sufficient for effectively mitigating phishing risk. Consider this scenario (adopted from Pillai et al 2012): a person A has more knowledge about phishing attacks than a person B, but B is more aware of the extent and limit of his/her knowledge. Knowing how much he/she knows and does not know enables B to adopt defensive strategies, such as contacting the sender of the email by phone, that help him/her better manage potential phishing threats. A, lulled into a false belief about his/her knowledge, may engage in a risky behavior (or missing information) following his/her belief. Confidence-accuracy miscalibration (Keren 1997), as A commits, is a common bias in human judgements and has been observed in the phishing literature (Hong et al 2013; Kumaraguru et al 2007).

Increasing the confidence-accuracy calibration (Alba and Hutchinson 2000; Keren 1997) is desirable for judgment under uncertainty but may be at the sacrifice of another important judgmental ability known as resolution (also called discrimination; Keren 1997; Yaniv et al 1991). Both calibration and resolution are fundamental aspects of a person's judgmental ability such as deception detection (Stone and Opel 2000). Calibration refers to the correspondence between a person's judgmental confidence and accuracy, while resolution, also known as discrimination, refers to the ability of a person to discern correct events from incorrect events (Bjorkman 1992; Yaniv et al 1991). Understanding how people differ in their calibration and resolution skills in phishing email detection and how to improve both skills are of critical importance to effectively combating phishing email attacks.

As the literature on calibration and resolution skills in phishing email detection is very scarce, we first provide an introduction of both skills. We then describe an empirical study to explore individual differences in calibration and resolution and the impacts of several factors on both skills, as follows.

Calibration and Resolution Skills in Phishing Email Detection

Conceptualization and measurement of calibration and resolution skills

Calibration and resolution skills are derived from judgmental confidence and accuracy. In judgment under uncertainty such as phishing email detection, a person's judgmental confidence, in addition to accuracy, is an important factor to consider (Keren 1997). Confidence represents a metaknowledge (Pillai et al 2012), or one's belief, that a specific statement, opinion, or decision is the most possible (Peterson and Pitz 1988). For example, confidence in one's phishing email detection represents one's metaknowledge of the ability to correctly discern phishing emails from genuine emails. As the behavioral outcomes of misjudging an email may not be observed immediately, judgmental confidence may be the direct drive of subsequent actions (Berger 1992).

Judgmental confidence is commonly measured as a subjective probability of an event (Juslin and Olsson 1997). For example, 100% means very confident, while 50% means by chance. Such a scale not only provides a more accurate measure of confidence than the Likert scale (Kumaraguru et al 2007) but is also consistent with the accuracy measure using percentage of correct answers (Bjorkman 1994), such as 100% correct or 50% correct. It was found that the half-range measure (i.e., 50-100%) is logically more reasonable and statistically more reliable than the alternative full-range measure (i.e., 0-100%) of confidence (Juslin and Olsson 1997). We therefore adopt the half-range measure in our study but cite the full-range measure when discussing prior literature, wherever applicable.

In a typical calibration research (e.g., Palmer et al 2013; Weber and Brewer 2004), a series (N) of judgmental tasks such as prediction questions are presented. For each question, the subject is asked to choose the correct answer and indicate the confidence in that answer. The simplest scenario is that each question has only two possible answers, e.g., "which river is longer, the Amazon or the Nile?" Such two-alternative questions dominate the calibration research and are very consistent with the phishing detection context since for each email we determine whether it is a genuine email or a phishing email, and how confident we are.

The measurement of confidence, denoted by f_n for the n-th judgment, is a probability score ranges from 50% to 100% with 100% indicating very certain and 50% indicating very uncertain. Accuracy of a judgment is measured based on its outcome (denoted by d_n for the n-th judgment), and in this study the outcome is whether a judgment on an email is correct or not. Following prior literature (Yates 1982), we set d_n to 1 for a correct judgment and 0 for an incorrect judgment. Metrics for measuring judgmental abilities such as the resolution skills are derived from the f_n and d_n measures. Table 1 lists the notations.

Notation	Explanation
N	Total number of judgments (i.e., number of questions)
n	The n-th judgement; n=1,2,…N
J	Number of judgmental categories; each category is defined by a confidence level or category assigned to a group of judgments; the half-range measure usually employs six categories: 50-59%, 60-69%, 70-79%, 80-89%, 90-99%, and 100%
j	The j-th judgmental category; j=1…J
N_j	The number of judgments in the j-th category; $\sum_{j=1}^{J} N_j = N$
f_n	A person's self-reported confidence in the n-th judgment, measured as a subjective probability score ranging from 50% to 100% (i.e., the half-range measure)
f_j	The confidence level in the j-th category, usually treated as the mean value of f_n in the category
d_n	Accuracy of the n-th judgment by a participant; d_n =1 for a correct judgment, and d_n =0 for an incorrect judgment
$\bar{d}$	Proportion of correct judgments of a participant, i.e., overall accuracy
$\bar{d}_j$	Mean accuracy in the j-th category

Table 1. Notations

The conceptualization and measurement of calibration and resolution skills are both derived from the Brier score (Brier, 1950), a common measure of how consistent one's judgment is with the actual events. The score, as shown in Equation (1), calculates the mean of the squared differences between confidence and accuracy in the series of N judgments:

$$\text{Brier score} = \frac{1}{N}\sum_{n=1}^{N}(f_n - d_n)^2 \quad (1)$$

The score ranges from 0 (when a person makes all correct judgments, i.e., d_n=1, n=1…N, with 100% confidence in each) to 1 (when a person misses all judgments, i.e., d_n=0, n=1…N, with 100% confidence in each), and .25 is earned by randomly choosing an answer in a judgment and assigning a confidence of 50%. Lower Brier score indicates better correspondence between confidence and accuracy.

Murphy (1973) provides a popular decomposition of the Brier score to extract more information (Yates 1982). It first groups the confidence measures (i.e., the f_n scores) into several (J) categories, and then calculates the squared differences between confidence and accuracy across the J categories (instead of across N individual judgments). Commonly six probabilistic categories are used, including 50-59%, 60-69%, 70-79%, 80-89%, 90-99%, and 100%; the categories are either used to directly measure judgmental confidence in the tasks, or derived from the confidence measures after the tasks. The decomposed Brier score (noted as Brier score') is shown in Equation (2):

$$\text{Brier score'} = \bar{d}(1-\bar{d}) + (\tfrac{1}{N})\sum_{j=1}^{J} N_j(f_j - \bar{d}_j)^2 - (\tfrac{1}{N})\sum_{j=1}^{J} N_j(\bar{d}_j - \bar{d})^2$$

$$= \text{Var(d)} + \text{CI} - \text{RI} \quad (2)$$

The three components in this decomposition provide detailed information about a person's judgmental ability: the first component is termed variance, uncertainty, or knowledge (Bjorkman 1992) in the judgments; the second component is called Calibration Index (or CI); and the third component is called Discrimination Index (DI) or Resolution Index (RI; Baranski and Petrusic 1994). CI is the measure of the calibration skills, which captures the correspondence between confidence and accuracy both within each category and among all the categories. It ranges between 0 (indicating perfect calibration, when $f_j = \bar{d}_j$ in each category) and 1 (indicating extreme miscalibration, when the subject feels 100% confident in each judgment but misses all), with a smaller value indicating a stronger judgmental ability. RI is the measure of the resolution skills, and it captures a person's ability to discern correct judgments from incorrect judgments, so that the higher RI is, the stronger judgmental ability the person exhibits. RI reaches maximum when $\bar{d}_j$ = 1 or 0, j=1…J; in other words, a person exhibits strongest resolution skills when he or

she assigns the correct judgments to the same confidence category (or level) and incorrect judgments to another confidence category (or level). Based on Equation (2), we have the following formulas for the calibration skills (or CI) and the resolution skills (or RI):

$$\text{CI} = (\tfrac{1}{N}) \sum_{j=1}^{J} N_j (f_j - \bar{d}_j)^2 \qquad (3)$$

$$\text{RI} = (\tfrac{1}{N}) \sum_{j=1}^{J} N_j (\bar{d}_j - \bar{d})^2 \qquad (4)$$

An illustration of the calibration and resolution skills

We adapt an example from Yaniv et al (1991) to illustrate the calibration and resolution skills in phishing email detection. Three persons (A, B, and C) make judgments of 50 emails; the email numbers, confidence levels, and accuracy results of the three persons are reported in Table 2. To be consistent with Yaniv et al (1991), we use the full-range (i.e., 0-100%) confidence measure in the example. For convenience, we list the emails in ascending order by the confidence levels and then descending order by the accuracy measures (i.e., accurate judgments first). For simplicity reason, we assume the three persons assign the same confidence to the same email judgment, although the specific accuracy values may differ (see Emails # 24, 25, 36, and 37). The metrics, including the total number of judgments (N), overall accuracy ($\bar{d}$), variance (Var(d)), CI, RI, and the decomposed Brier score are also listed in the table.

Email #	Confidence (f_n)	Accuracy (d_n)		
		A	B	C
1	10%	1	1	1
2	10%	1	1	1
3	10%	1	1	1
4	10%	1	1	1
5	10%	0	0	0
6	10%	0	0	0
7	10%	0	0	0
8	10%	0	0	0
9	10%	0	0	0
10	10%	0	0	0
11	10%	0	0	0
12	10%	0	0	0
13	10%	0	0	0
14	10%	0	0	0
15	10%	0	0	0
16	10%	0	0	0
17	10%	0	0	0
18	10%	0	0	0
19	10%	0	0	0
20	10%	0	0	0
21	40%	1	1	1
22	40%	1	1	1
23	40%	1	1	1
24	40%	1	1	0
25	40%	1	0	0
26	40%	0	0	0
27	40%	0	0	0
28	40%	0	0	0
29	40%	0	0	0
30	40%	0	0	0
31	60%	1	1	1
32	60%	1	1	1
33	60%	1	1	1
34	60%	1	1	1
35	60%	1	1	1
36	60%	0	1	1
37	60%	0	0	1
38	60%	0	0	0
39	60%	0	0	0
40	60%	0	0	0
41	80%	1	1	1
42	80%	1	1	1
43	80%	1	1	1
44	80%	0	0	0
45	80%	0	0	0
46	100%	1	1	1
47	100%	1	1	1
48	100%	1	1	1
49	100%	1	1	1
50	100%	0	0	0
	N	50	50	50
	$\bar{d}$	0.420	0.420	0.420
	Var(d)	0.244	0.244	0.244
	CI	0.016	0.012	0.016
	RI	0.040	0.044	0.056
	Brier score'	0.220	0.212	0.204

Table 2. An illustration of judgmental skills

It can be seen that all three persons make the same percentages of correct judgments ($\bar{d}$ = 42%), so that the variances or uncertainties in their judgments are the same (Var(d) = 0.244). This suggests that

detection accuracy alone is not sufficient to compare the detecting abilities of the three persons, and we need to consider their confidence as well. In terms of the calibration skills, B has a lower CI (.012) than A and C (.016 for both) due to the fact that B's judgments are better aligned at the 40% and 60% confidence levels: at the 40% level, B's accuracy is 40% (perfectly calibrated in this category) as compared to 50% for A and 30% for C; at the 60% level, B's accuracy is 60% (again, perfectly calibrated in this category) as compared to 50% for A and 70% for C; in neither of the categories is A or C calibrated. From this perspective, B seems to do a better job than A and C in making the judgments.

Nevertheless, in terms of RI, C exhibits stronger judgmental ability than both A and B, as C has the highest RI score. This happens because C's judgmental accuracy (30%) at the 40% confidence level is less than the accuracy of both A (50%) and B (40%), but at the 60% confidence level, C's judgmental accuracy (70%) is greater than that of A (50%) and B (60%). As 60% confidence is stronger than 40% confidence, this suggests that C is more aware of his/her ability of making correct judgments at a higher confidence level and the inability of making correct judgments at the lower confidence level than both A and B. A and B, nonetheless, are less aware of their potentials and limits of making correct/incorrect judgments. Especially for A, his/her judgmental accuracy (50%) is the same across the 40% and 60% confidence levels, suggesting that A, who has the lowest RI score, is least capable of discerning correct judgments from incorrect ones.

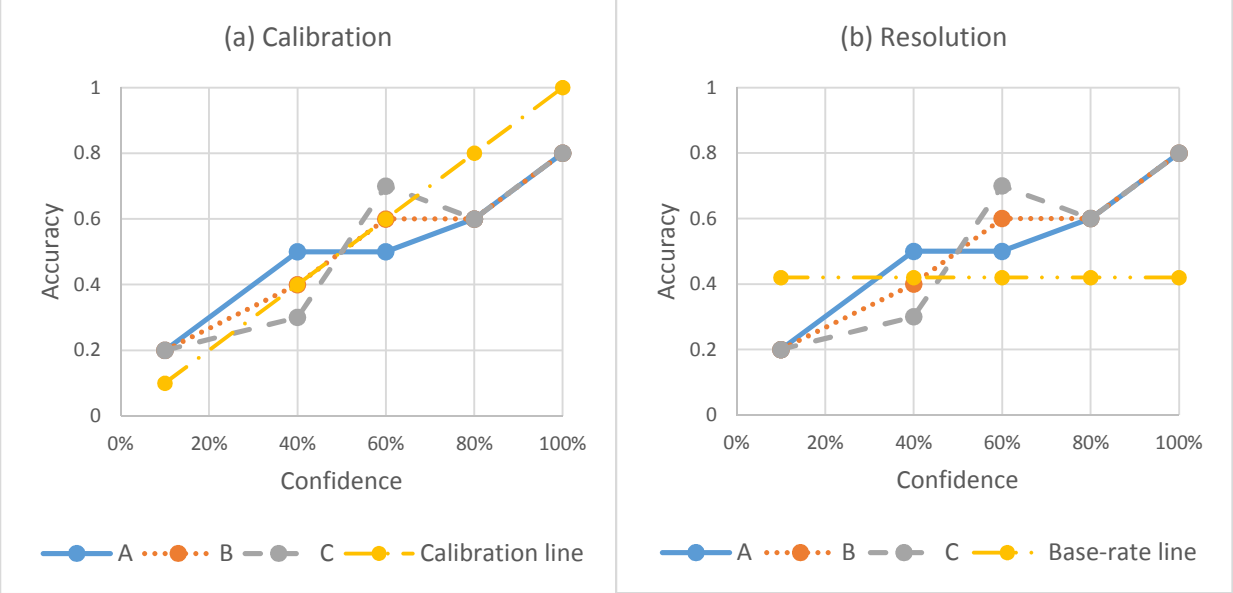

Figure 1 Calibration versus Resolution

We use Figure 1, adapted from Yaniv et al (1991) and based on Table 2, to compare the two types of skills. For calibration, the aim is to ensure that the accuracy and confidence measures are consistent in each of the judgmental categories, so that the calibration curve should be consistent with the 45° calibration line. Apparently, the calibration curve of B (see Figure 1-a) is closer to the calibration line, exhibiting stronger calibration skills of B than both A and C. The aim of resolution, on the other hand, is to "differentiate instances when an event is going to occur from those when it is not (Yaniv et al 1991, p. 612)." In other words, the resolution skill aims to assign judgments with the same outcome (correct or not) to the same confidence category, or "maximum resolution occurs if correct responses and incorrect responses are sorted into different categories (Bjorkman 1994, p. 4)." On Figure 1-b, it is reflected in the vertical displacement of each $\bar{d}_j$ from the base-rate line (i.e., the $\bar{d}$ line): the larger displacement, the higher RI. In fact, C exhibits larger displacements in the 40% and 60% judgmental categories, therefore having a higher RI; this is achieved, however, at the cost of the calibrations in both categories as compared to B.

It should be noted that the Brier score' in Table 2 suggests that C is overall better than A and B in judging the emails. This is due to the strong resolution skills of C in the current case. Nevertheless, it doesn't mean that the RI is always consistent with the Brier score, as in other scenarios the CI may be a major contributor of the score. In other words, we need to examine both to fully understand a person's judgmental abilities in dealing with phishing emails.

Influential factors of calibration and resolution skills in phishing email detection

The above illustration shows that calibration and resolution skills are distinct aspects of judgmental abilities and may have different implications for research and practice. As prior literature shows that different factors, such as performance feedback and environmental feedback in training (Stone and Opel 2000), have different effects on both, we thus conduct a study to explore the effects of potential factors on both skills in phishing email detection.

While much research has been done to examine calibration skills in various task environments, limited progress has been made to examine resolution skills despite the common belief that both are equally important in judgment (Keren 1997). Therefore, to find a list of potential influential factors, we searched the phishing detection literature and also compared to the calibration and resolution literature. The process helped generate a short list of factors that we can manage in this study, including: individual characteristics such as one's online transaction experience, prior victimization of phishing attack, self-efficacy belief, number of emails received each day, gender, age, and education, and task factors such as familiarity with the business entity in an email, familiarity with the email received, time spent to judge the email, variability of time spent to judge all the emails, and overall easiness/hardness of the judgments. The individual characteristics specify how a person differs from others in the general ability of identifying phishing emails. Some of the factors, for instance, were recognized from the study on phishing vulnerability by Vishwanath et al (2011), including email load, knowledge of phishing, and computer self-efficacy. Other demographic factors were recognized from Wang et al (2012). The task factors specify how the particular judgmental tasks, such as the difficulties of judging the emails (as some phishing emails are more deceptive than others) and personal familiarity with the emails or email senders, may influence the calibration and resolution skills. As most of the above factors were examined for their impacts on judgmental accuracy (either vulnerability or response likelihood) only, and few were examined in the calibration and resolution literature, we therefore conduct an exploratory study on these factors to investigate their different impacts on the judgmental skills. Specifically, a factor improves calibration skills if it reduces the CI, and it improves resolutions skills if it increases RI. The research method and results are described as follows.

Research Method and Results

Research design

We conducted an online experiment to measure individuals' calibration and resolution skills in phishing email detection and to empirically test the effects of some potential factors. This study is a part of a larger research project that investigates individuals' phishing detection abilities. In this experiment, each participant made judgments of 16 emails randomly chosen from a pool of 50 phishing emails and genuine business emails. Their judgmental confidence, accuracy as well as the influential factors were measured with either self-reported questionnaire or objective data retrieved from computer logs. In the experiment, the participants first completed items measuring phishing detection self-efficacy belief, and then finished the email judgment tasks in which sixteen email images were presented sequentially and the participants were asked to judge whether each email was legitimate or not. The participants were also asked to indicate how confident they were for each judgment (with a number between 50 and 100), whether they had received or seen the email before, and how familiar they were with the business entity indicated in the email. At the end of the experiment, the participants completed other questions measuring their genders, ages, educations, numbers of daily emails received, online transaction experience, and prior victimization of phishing attack.

Measurement methods

Table 3 shows the measurements of some of the independent variables. Other independent variables such as age, gender, and education were each measured with a single item. As the unit of analysis is individual, we derived mean values of the familiarity measures for business entities and emails across the 16 emails for each subject, which capture the subject's overall familiarity with the entities or emails. Online transaction experience, prior victimization of phishing attack, and perceived self-efficacy in phishing detection were each measured with the sum values of their corresponding items. Time to judgment was recorded online between showing the email image and the click the participant made for the judgment. The total time spent on the sixteen emails was then calculated to measure time to judgment. Following prior literature in behavioral decision making (Brucks 1985; Payne 1976; Stone 1994), the coefficient of variation (CV) of the time spent on each email was calculated to measure variability of time in judgments. Finally, easiness of the tasks were derived from the judgmental results of the emails, as Table 3 shows.

Independent variables	Measurement methods
Online transaction experience	1) buying products or services online with a credit card, a debit card, or a payment service such as PayPal. 2) accessing bank accounts (such as checking, saving, mortgage) online. 3) paying bills (such as electronic, utility, credit cards, loans) online. 4) buying and selling stocks or mutual funds online.
Prior victimization of phishing attack	1) someone used or attempted to use your credit cards without permission. 2) someone used or attempted to use your accounts such as your wireless phone account, bank account or debit/check cards without your permission. 3) someone used or attempted to use your personal information without permission to obtain new credit cards or loans, run up debts, open other accounts, or commit other frauds.
Self-efficacy belief in phishing detection	1) I can recognize phishing emails. 2) I can differentiate phishing emails from legitimate ones.
Number of emails received each day	Roughly on average how many emails do you receive per day?
Familiarity with the business entity	How familiar do you think you are with the business entity indicated in the email?
Familiarity with the email received	Have you personally received or seen this particular email before this survey?
Time to judgment	Recorded online
Variability of time in judgments	The coefficient of variation (CV) of the time spent on the emails was calculated
Easiness of the tasks	The mean value of the easiness of judging the emails that a subject received, where the easiness of each email was derived from the proportion of subjects who judged the same email correctly

Table 3 Measurement of independent variables

The measurement of the dependent variables (i.e., CI and RI) follows Equations (3) and (4). Specifically, the measures of confidence and accuracy of the 16 emails were grouped into six categories (i.e., 50-59%, 60-69%, etc.), and the mean accuracy ($\bar{d}_j$) and mean confidence (f_j) in each category were calculated, from which CI and RI were derived.

Data analysis and results

Of the 600 subjects who participated in the study, 8 failed to follow the instruction and were excluded from the sample, resulting in 592 valid observations. We first conduct a descriptive analysis on the calibration and resolution skills of the subjects and compare to other metrics. As shown in Table 4, miscalibration is common in the subjects, as the mean confidence (0.8089) is much higher than the mean accuracy (0.6670). Var(d) is a major contributor to the Brier score', and both RI and CI are much smaller. Nevertheless, Baranski and Petrusic (1994) point out that the calibration score and the resolution score

above .10 are rarely encountered, suggesting that our results are normal. But still, the relative small RI, as compared to Var(d), means that a large amount of uncertainties in phishing detection cannot be accounted for by the resolution skills. Due to page limitation, the correlation matrix of the variables is not provided but is available upon request from the authors.

	Accuracy	Confidence	Var(d)	CI	RI	Brier score'
Mean	0.6670	0.8089	0.1995	0.0906	0.0414	0.2487
Std. Dev.	0.1504	0.1213	0.0509	0.0801	0.0339	0.1063
C.I. (5%)	0.3750	0.5658	0.1094	0.0084	0	0.0994
C.I. (95%)	0.8750	0.9744	0.2500	0.2585	0.1085	0.4709

C.I. – Confidence Internal; Sample size = 592

Table 4 Descriptive information of the metrics

We ran Multiple Liner Regressions (MLR) to test the impacts of the influential factors on CI and RI. Variance Inflation Factors (VIFs) were included in the regressions to detect potential multicollinearity, and no significant issues were noticed. The results are reported in Table 5. Interestingly, all the significant predictors of CI are task factors, including familiarity with emails, time to judgment, variability of time in judgments, and easiness of the tasks. Specifically, both time to judgment and easiness of the tasks reduce CI and therefore improve calibration skills. In other words, the more time a person spent judging the emails, the more calibrated his/her confidence and accuracy are, which is consistent with the calibration literature (Palmer et al 2013; Weber and Brewer 2004). For the easiness of tasks, similarly, the easier the tasks are, the more calibrated a person is, which is also consistent with the hard-easy effect widely recognized in the calibration literature, which suggests that harder tasks tend to yield miscalibration in subjects (Juslin and Olsson 1997; Keren 1997). Both familiarity with emails and variability of time increase CI and therefore reduce calibration skills, which are also evidenced in calibration literature: familiarity with an email makes a person relaxed and less alert to potential risks, while variations in time to judgment result in selective retrieval of supporting evidence and neglect of disconfirming information, contributing to miscalibration in judgment (Alba and Hutchinson 2000). None of the individual differential factors are significant in this study.

Independent variable	CI			RI		
	β	t-value	p-value	β	t-value	p-value
Online transaction experience	0.072	1.658	0.098	0.150	3.383	0.001***
Prior victimization	-0.045	-1.096	0.273	0.100	2.393	0.017*
Self-efficacy of phishing detection	0.021	0.504	0.614	-0.022	-0.536	0.592
No. of emails received each day	0.000	-0.005	0.996	0.043	1.066	0.287
Gender	-0.054	-1.318	0.188	0.049	1.189	0.235
Age	0.004	0.086	0.932	-0.082	-1.920	0.055
Education	0.020	0.484	0.629	-0.010	-0.244	0.807
Familiarity with entity	0.055	1.210	0.227	-0.100	-2.167	0.031*
Familiarity with email	0.099	2.294	0.022*	0.048	1.086	0.278
Time to judgment	-0.130	-3.074	0.002**	0.078	1.816	0.070
Variability of time	0.158	3.672	0.000***	-0.130	-2.987	0.003**
Easiness of the tasks	-0.120	-2.993	0.003**	-0.037	-0.901	0.368
R^2	.085			.057		
Adjusted R^2	.066			.037		

* - significant at .05 level; ** - significant at .01 level; *** - significant at .001 level

Table 5 Results of Multiple Liner Regressions

For RI, the results show four significant predictors as well, including online transaction experience, prior victimization of phishing attacks, familiarity with email entity, and variability of time in judgment. Specifically, both online transaction experience and prior victimization of phishing attacks increase RI and therefore improve resolution skills, which is consistent with prior literature on detection accuracy (Downs et al 2007). Both familiarity with email entity and variability of time in judgment decrease RI and therefore reduce resolution skills, which is also consistent with prior literature on judgmental abilities (Alba and Hutchinson 2000). The study therefore confirms that different factors have different effects on the calibration and resolution skills.

Concluding Remarks

In this study, we examine the calibration and resolutions skills in phishing email detection and conduct an empirical test to explore the impacts of some task factors and personal characteristics on both skills. Based on an illustration (in Table 2 and Figure 1), we show that the calibration and resolution skills are two distinct sets of abilities in judging phishing emails, and that there is an intrinsic conflict between the two (Keren 1997) that should be seriously addressed in order to design appropriate mechanisms (such as training and feedback) to improve both. The empirical test shows that both skills respond to different sets of factors (except for variability of time, which influences both), providing further support to the literature (e.g., Stone and Opel 2000).

The contributions of the study are two-fold. For one, we illustrate the importance of studying both calibration skills and resolution skills in phishing email detection, calling for more research on these two components of judgmental abilities. For two, our study is the first to empirically test both skills in phishing email detection and compare them side by side to show their differences. The different impacts of influential factors on calibration and resolution skills have theoretical and practical implications.

Theoretically, our findings suggest that only task-related factors such as familiarity with the emails to be judged, time to judgment, variability of time, and overall task easiness may influence calibration skills, but none of the individual characteristics such as prior experience with phishing attacks or self-efficacy in phishing detection has an impact. This implies that the process how a person handles emails (e.g., how much time is spent on examining the email) is a main determinant of the calibration skills, and such skills can be improved by moderating the task-related factors recognized above. For instance, a person can be trained to effectively allocate time to examine emails in his/her inbox rather than selectively examining emails which may result in neglect of important security cues. One the other hand, our findings show that both individual characteristics and task-related factors influence the resolution skills, implying that both types of factors should be considered when designing mechanisms to improve the skill. For instance, when a person learns knowledge about the risks of online transactions and phishing attacks, he or she should also learn to apply the knowledge appropriately or adequately (with sufficient time, for instance) when judging the emails. The distinctive impacts of individual and task factors should be further investigated.

Practically, our study recognizes a list of factors that should be focused on when training individuals to improve their phishing email detection abilities. The individual characteristics demand that general knowledge about online transactions and risks of phishing attacks should be taught to individuals, and the task factors require that the individual be trained to allocate time (or other cognitive resources) appropriately to make better judgments of phasing emails.

The study has a number of limitations. First, only a small number of influential factors were examined, so that the variances explained in both CI and RI are relatively small. In the future, other factors should be investigated for their impact on the calibration and resolution skills. Second, the experimental setting may influence a person's judgments as the subjects are aware of the objective of the study. In the future, other methods may be applied to validate the generalizability of the findings of this study.

References

Alba, J. W., and Hutchinson, J. W. 2000. "Knowledge Calibration: What Consumers Know and What They Think They Know," Journal of Consumer Research (27:2), pp.123–156.

Baranski, J. V. and Petrusic, W. M. 1994. "The calibration and resolution of confidence in perceptual judgments," Perception and Psychophysics (55), pp. 412-428.

Berger, I.E. 1992. "The nature of attitude accessibility and attitude confidence: a triangulated experiment," Journal of Consumer Psychology 1(2) pp. 103-123.

Bjorkman, M. 1992. "Knowledge, calibration, and resolution: a linear model," Organizational Behavior and Human Decision Processes (51), pp. 1-21.

Bjorkman, M. 1994. "Internal cue theory: calibration and resolution of confidence in general knowledge," Organizational Behavior and Human Decision Processes (58), pp. 386-405.

Brier, G. W. 1950. "Verification of forecasts expressed in terms of probability," Monthly Weather Review (78), 1-3.

Brucks, M. 1985. "The effects of product class knowledge on information search behavior," Journal of Consumer Research (12:1), pp. 1–16.

DePaulo, B.M., Charlton, K., Cooper, H., Lindsay, J.J. and Muhlenbruck, L. 1997. "The Accuracy-Confidence Correlation in the Detection of Deception," Personality and Social Psychology Review (1:4), pp. 346-357.

Eastin, M. S., and LaRose, R. 2000. "Internet Self-Efficacy and the Psychology of the Digital Divide," Journal of Computer-Mediated Communication 16(1).

Gigerenzer, G., Hoffrage, U. and Kleinbolting, H. 1991. "Probabilistic mental models: A Brunswikian theory of confidence," Psychological Review (98), pp. 506-528.

Heath, C., and Tversky, A. 1991. "Preference and belief: Ambiguity and competence in choice under uncertainty," Journal of Risk and Uncertainty 4(1), pp. 5–28.

Hong, K.W., Kelley, C.M., Tembe, R., Murphy-Hill, E., and Mayhorn, C.B. 2013. "Keeping Up With The Joneses: Assessing Phishing Susceptibility In An Email Task," Meeting of the Human Factors and Ergonomics Society.

Juslin, P., and Olsson, H. 1997. "Thurstonian and Brunswikian Origins of Uncertainty in Judgment: A Sampling Model of Confidence in Sensory Discrimination," Psychological Review (104:2), pp. 344–366.

Keren, G. 1997. "On the calibration of probability judgments: Some critical comments and alternative perspectives," Journal of Behavioral Decision Making (10), pp. 269-278.

Kumaraguru, P., Rhee, Y., Sheng, S., Hasan, S., Acquisti, A., Cranor, L. and Hong, J. 2007. "Getting Users to Pay Attention to Anti-Phishing Education: Evaluation of Retention and Transfer," APWG eCrime Researchers Summit, October, 4-5, 2007, Pittsburgh, PA, USA.

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

Li, Y. 2013. "The impact of disposition to privacy, website reputation and website familiarity on information privacy concerns," Decision Support Systems (57), pp. 343–354.

Liberman, V. and Tversky, A. 1993. "On the Evaluation of Probability Judgments: Calibration, Resolution, and Monotonicity," Psychological Bulletin (114:1), pp.162-173.

Murphy, A. H. 1973. "A new vector partition of the probability score," Journal of Applied Meteorology (12), pp. 595-600.

Palmer, M.A., Brewer, N., Weber, N. and Nagesh, A. 2013. "The Confidence-Accuracy Relationship for Eyewitness Identification Decisions: Effects of Exposure Duration, Retention Interval, and Divided Attention," Journal of Experimental Psychology: Applied (19:1), pp. 55–71.

Payne, J. W. 1976. "Task complexity and contingent processing in decision making: An information search and protocol analysis," Organizational Behavior and Human Performance 16(2), pp. 366–387.

Peterson, D. K., and Pitz, G. F. 1988. "Confidence, uncertainty, and the use of information," Journal of Experimental Psychology: Learning, Memory, and Cognition (14:1), pp. 85–92.

Stone, D. N. 1994. "Overconfidence in Initial Self-Efficacy Judgments: Effects on Decision Processes and Performance," Organizational Behavior and Human Decision Processes (59:3), pp. 452–474.

Vancouver, J. B., Thompson, C. M., Tischner, E. C., and Putka, D. J. 2002. "Two studies examining the negative effect of self-efficacy on performance," Journal of Applied Psychology (87:3), pp. 506–516.

Vishwanath, A., Herath, T., Chen, R., Wang, J., and Rao, H. R. 2011. "Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model," Decision Support Systems (51:3), pp. 576–586.

Wang, J., Herath, T., Chen, R., Vishwanath, A., and Rao, H. R. 2012. "Phishing Susceptibility: An Investigation Into the Processing of a Targeted Spear Phishing Email," IEEE Transactions on Professional Communication (55:4), pp. 345–362.

Weber, N., and Brewer, N. 2004. "Confidence–Accuracy Calibration in Absolute and Relative Face Recognition Judgments," Journal of Experimental Psychology: Applied (10:3), pp. 156–172.

Wright, R. T., and Marett, K. 2010. "The Influence of Experiential and Dispositional Factors in Phishing: An Empirical Investigation of the Deceived," Journal of Management Information Systems (27:1), pp. 273–303.

Yaniv, I., Yates, J.F., and Smith, J.E.K. 1991. "Measures of discrimination skill in probabilistic judgment," Psychological Bulletin (110:3), pp. 611-617.

Yates, J. F. 1982. "External correspondence: Decomposition of the mean probability score," Organizational Behavior and Human Performance (30), pp. 132-156.