Who Provides Phishing Training? Facts, Stories, and People Like Me

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

Humans represent one of the most persistent vulnerabilities in many computing system. Since human users are independent agents who make their own choices, closing these vulnerabilities means persuading users to make different choices. Focusing on one specific human choice – clicking on a link in a phishing email – we conducted an experiment to identify better ways to train users to make more secure decisions. We compared traditional facts-and-advice training against training that uses a simple story to convey the same lessons. We found a surprising interaction effect: facts-and-advice training works better than not training users, but only when presented by a security expert. Stories don't work quite as well as facts-and-advice, but work much better when told by a peer. This suggests that the perceived origin of training materials can have a surprisingly large effect on security outcomes.

ACM Classification Keywords

K.6.5. Security and Protection: Miscellaneous

Author Keywords

phishing; training; stories; security

INTRODUCTION

Modern computing systems are extremely complex, and often involve hundreds or thousands of separate components, chips, applications, and people operating together to achieve their desired function. All systems that are this complex have bugs and problems – instances when the system does not function as intended. Sometimes, those bugs can be exploited by creative people to cause the system to operate on their behalf, rather than on the behalf of the owner and user. We call these bugs "vulnerabilities", and much of the challenge in using modern computing systems involves finding and eliminating vulnerabilities in the system.

However, computing systems do not function independently of their human users; indeed a computing system that never interfaces with a human largely has no purpose. Humans are integral parts of modern computing systems: they provide inputs and read outputs of the system; they design, build, program, and configure the systems to work on their behalf; and they make a variety of critical decisions that the computers are unable to make themselves. In these roles, humans function as a critical component of modern computing systems, and as such, they can also create vulnerabilities in the system.

However, unlike computers, humans cannot be programmed to perform. They cannot easily be patched to change their behavior when a vulnerability is discovered. And they do not behave deterministically, so even if they behave correctly today, they might make a different decision tomorrow. "Patching vulnerabilities" in the human users of systems is one of the most challenging aspects of computer security. Since humans are independent agents that make their own decisions, they must be *persuaded* to want to change their behavior. And most human users are not experts at using computers, and therefore often need to be *trained* to learn how to make more secure decisions. These two aspects – end user motivation and end user training – make vulnerabilities introduced by humans particularly difficult to address.

Currently, the most common method of changing end-user security behaviors is *security education and training* – mostly end-user education campaigns consisting of webpages, flyers and posters, along with basic computer literacy training in schools. This training leaves large numbers of people vulnerable. We seek to identify new methods that can be used to "patch vulnerabilities" in the human components of computing systems by motivating users to make more secure decisions and helping users understand their decisions.

Focusing on one specific human vulnerability – phishing emails – we conducted a field experiment to test better methods to train users. We tested two very different types of training messages: traditional facts-and-advice training, and narrative stories about other people who have previously had phishing problems. We also tested if these messages would be received differently if they appeared to come from a security expert or from a person like them. We found an interaction effect: facts-and-advice are more effective when provided by security experts, but stories are more effective when told by other people like them. This interaction indicates that a new potential avenue for security training (stories from peers) may be effective, and that previous research that used narratives [26] may be underestimating their effectiveness because their narratives came from experts.

CHI'18, April 21–26, 2018, Montreal, Canada

ISBN 123-4567-24-567/08/06. . . $15.00
DOI: http://dx.doi.org/10.475/123_4

LITERATURE REVIEW

Human Vulnerabilities

Human vulnerabilities are a very widely exploited method of breaking into and misusing computing systems, and attacks that exploit these vulnerabilities are a large source of costs in many organizations. Phishing scams alone cost organizations approximately $3.2 billion in 2007 [18]. Many companies spend as much as $3.7 million a year preventing phishing and spear phishing attacks [13]. Nelson (based on a study by Cloudmark) estimated spear phishing attacks can cost an estimated $1.6 million per incident [27].

Trying to address these human vulnerabilities is big business. Many organizations, including the US Federal Bureau of Investigation, the US Computer Emergency Response Team, and most large technology companies have a large numbers of webpages and training campaigns that are designed to train the general public in how to protect themselves and their computers from cyber attack [32]. There are a large number of companies that sell software and training solutions to organizations that can be used internally to train their employees in cyber security.

However, despite all of this effort to train end users, a large number of end users are still vulnerable. Indeed, at a recent public talk, Rob Joyce, the chief of Tailored Access Operations at the US National Security Agency (effectively the chief hacker for the US) discussed how exploiting technical vulnerabilities is effective when they are available, but the most reliable method of breaking into computers is to get their human users to make a bad decision [20]. That is, he claimed that human vulnerabilities provide a more reliable and consistent method for gaining unauthorized access to computing systems than technical vulnerabilities. And he emphasized that this is true not only for home computer users, who commonly lack the technical expertise to secure their computers [38], but it is also true at the level of large organizations and nation-states that spend large amounts on computer security.

Phishing

We focus on a specific, known vulnerability in human users: phishing scams. A phishing scam involves sending a user an email that is pretending to be from a person or organization that the user knows, and asking that user to take some action that seems like it would benefit the user or organization, but really benefits the attacker. The most common example of a phishing scam involves the attacker sending an email asking the user to click on a link and enter in their username and password on the resulting webpage (which is often a fake webpage controlled by the attacker but designed to mimic the appearance of a webpage for the trusted organization). Other examples of phishing emails include fake emails asking users to click on a link to install or run a new application, or asking users to click on an insecure link that ends up executing some other type of technical attack (e.g. drive-by download of ransomware).

Phishing messages are very complicated to identify because they are often intentionally designed to mimic real, trustworthy messages [5]. There are two major types of phishing messages that are commonly studied: phishing websites that mimic real, known websites but surreptitiously collect usernames and passwords or provide malicious downloads; and phishing emails that mimic real emails but contain links to phishing websites where data is collected or distributed. This pattern is quite common – a first message (email) that directs vulnerable users to a second message (website) where the actual bad things happen. In addition to email, phishing has been found on voicemail (vishing) and text chat (smishing) and others [33].

Phishing messages are difficult for users to detect. Schechter et al. found that most security indicators on emails and websites are generally small and difficult to notice. They found that it is difficult for end users to notice the absence of an indicator than the presence of an indicator [35]. Phishing messages often take advantage of this by including fake versions of common security indicators [11]. Phishing messages that appear to come from friends or other trusted sources are also not scrutinized as closely and users are more likely to click on such messages [19]. Messages on Facebook are more likely to be clicked than messages via email, as are messages that invoke curiosity or fit the recipient's expectations [3].

Phishing Training

A number of researchers have attempted to design interventions that train, educate, and persuade end users. Table 1 contains a comparison of a number of these phishing studies.

Kumaraguru et al. [26] conducted a series of phishing studies intended to teach end users how to identify common phishing messages, and then measure behaviorally whether they clicked on links in (fake) phishing messages. Their major finding involves the timing of training: training messages that are provided to users in real time when they click on an inappropriate link – which they called "embedded training" – are significantly more effective than similar training messages that can be perused at the user's leisure [23]. Embedded training provides a strong motivation to learn – the user just made a mistake – and also provides fast and effective feedback to users at the time they are most receptive to it [26]. They also found that using a comic where characters explain phishing concepts to teach the phishing lesson was more effective than traditional webpages with text [24].

While their original studies used role playing in a research lab, they were able to validate these findings in two real-world organizations [25, 22]. They were able to reduce click rates from an initial 40%-50% down to rates closer to 20%. However, this still leaves a large number of people vulnerable to phishing attacks. Also, an better controlled experiment that replicated the embedded training + comic structure with longer periods of time between training and testing found no improvement in click rate [5].

Another approach to training focuses specifically on the URLs in the phishing links. "Anti-phishing Phil" is a web-based game designed by a team at Carnegie Mellon that trains users to identify phishing URLs. It uses a fun game framed around actual fishing (URLs are hooks catching your fish) to educate

Authors	Publication	Location	N	Click	Outcome	Repeat
Ferguson [12]	EDUCAUSE 2005	Field	512	80%	Click	None
Wu et al. [39]	CHI 2006	Lab	30	52%	Click	None
Jagatic et al. [19]	CACM 2007	Field	921	72% / 16%	Info	None
Kumaraguru et al. [23]	CHI 2007	Lab	30	90%	Click	None
Kumaraguru et al. [24]	eCrime 2007	Lab	42	90%	Click	7 days
Kumaraguru et al. [25]	eCrime 2008	Field	311	42%/39%	Both	2 and 7 days
Kumaraguru et al. [22]	SOUPS 2009	Field	515	52%/51%/45%	Both	28 days
Caputo et al. [5]	IEEE S&P Mag, 2014	Field	1,359	60%	Click	7 months
This Paper		Field	1,945	11.7%	Click	2,7 and 42 days

Table 1. Summary of previous research about Phishing clicking and training. Field studies were conducted with unsuspecting subjects, and thus likely represent more accurate estimates of click rates. Click rates in this table are before subjects received training (if the study included training); if multiple rates are listed, the original paper reported separate rates for different conditions. Most studies considered a 'click' to be falling for the scam, but a few considered entering personal information ('info') into the subsequent webpage.

users how to look at a URL and identify which website it actually goes to [36].

Many organizations in the information security industry have also adopted the "embedded training" approach to phishing prevention. A number of companies, including Wombat Security, PhishMe, and Symantec, sell vended solutions that allow an organization to send test phishing messages to employees and provide embedded training when users click on those messages. This approach has become an industry standard way of protecting an organization against phishing attacks. However, there are times when this approach can backfire and actually increase how frequently users click on phishing messages. This can happen, for example, if users know about the training and expect to receive actually-benign phishing messages that they then click on specifically to see the newest training pages [16].

While the idea of embedded training (displaying training after clicking on a fake phishing message) has become an industry standard, there is still much work to be done to maximize the effectiveness of the training messages themselves. Most organizations have "help" webpages that are intended to help users learn how to deal with phishing. Rader and Wash [32] catalogued a number of these websites from many different vendors and analyzed them. The vast majority of these webpages fit a common pattern: 1) they first provide decontextualized factual information such as a definition of phishing and generic examples of harm; and 2) they provide generic advice in the form of 2nd person imperative statements ("you should do X" or "don't do X"). This facts-and-advice structure is useful to users, but is often incomplete from a user's point of view. It often does not contain information about who might be attacking, or the detailed social and personal consequences of actual attacks [32].

Stories

Recent work has suggested that these help webpages are not the only source of security information for end users. Rader et al. [31] found that many end users learn about security by talking to each other, and specifically by telling stories about incidents to each other. They catalogued a number of these stories, and found that such stories are often told by family and friends about specific security incidents, and often are told as a way of conveying "lessons" about security. Rader and Wash [32] re-analyzed these stories to find that, relative to help webpages or news articles, stories tend to focus disproportionately on attackers and their motivations. Das et al. [9] conducted an interview study and also found significant social influences on security behaviors, often as a way to "warn" others about specific, observable threats.

Stories are narratives in which characters (people) experience predicaments (problems, context) [21]. They are told as a sequence of events connected together to establish a plot. People in the stories have intentions (goals) and take actions. In response to those actions, other things happen (cause and effect). Framing knowledge as a sequence of actions rather than as factual statements more closely resembles how people learn from experience [21].

Human beings very naturally tell stories about real-life events to each other. They use stories as a way to learn from others' experiences so they can react appropriately in similar situations. For example, the purpose of gossip is not to spread factual knowledge about people, but instead is used to convey lessons about how to behave in various circumstances [2]. We suspect that stories can be intentionally used to train users about security in general, and phishing in particular.

METHODS

In this project, we focus on how to design appropriate training messages that will help users avoid clicking on phishing emails. Existing literature suggests that many users learn about computer security issues through social influences – that is, they learn about security incidents, problems, and solutions by talking to other people and hearing stories and warnings from them [30, 9]. We wanted to know if this method of learning can be used to improve security training messages in organizations. To that end, we created a phishing experiment: we sent a large number of users multiple phishing emails with randomly assigned embedded training. When any user clicks on the link in any of the phishing emails, they receive a training message. We randomly assigned training messages so we could identify the causal effects of the different forms of training. And then we measured future click rates when subsequent phishing emails were sent to the same users.

We created five different versions of our training message. First, we wanted to see if social stories might provide a more effective training message than more traditional facts-and-advice messages. However, we noted that the existing literature found an important difference for learning from social stories: the stories or warning almost always came from friends, family, colleagues, or other trusted, known individuals, whereas the facts-and-advice are always written from the perspective of the experts working for the organization [32]. This known individual vs organization's expert distinction might be important in the way that end users interpret the messages.

Therefore, we created messages in a 2x2 structure. Across one dimension we varied whether the training message contained a facts-and-advice message, or whether it contained a story about a previous phishing incident. Across another dimension, we varied who was seen as communicating the information: either an expert working for the organization, or a generic person similar to the end user. Finally, none of these four conditions provide a strong control condition to evaluate the effectiveness of these messages against. So we includes a fifth condition with no message as a control condition: do any of these messages perform significantly better than not providing training at all?

Phishing Campaign

To test these messages, we conducted a phishing campaign at our university. Our university is a large flagship state university in the midwestern United States. We designed our study to target a randomly chosen set of 2000 staff members at the university to each receive a series of four phishing emails spread out over the period of two months. Each of these phishing messages would contain a personalized link that would allow us to track whether the staff member clicked on each message. The first email where the link is clicked will take the staff member to a training message page with the training message for the experimental condition that this person was randomly assigned to; clicks in subsequent emails simply went to a blank webpage. This study and all pre-studies were approved by our institution's IRB.

Subject Selection

Working with the university central IT Services organization, we identified a list of approximately 7,500 staff members at the university who have email accounts with the university. This includes a wide variety of different types of jobs, including academic staff (e.g. secretaries who work in department or administrator offices), facilities staff (groundskeepers, HVAC repairmen), medical and veterinary staff who work with university-affiliated clinics, athletics staff, accounting staff, development and advancement staff, food service staff, museum staff, and many other staff members associated with specific activities conducted by the university. We excluded from this list all individuals who are also faculty at the university (which removes deans, department chairs, and most senior administrators), anyone listed as working in an "IT Services" or information technology organization, and individuals with special protections on their email account (such as the head football coach). While everyone in this population works for the large university, we believe that the wide range of job titles illustrates that this population represents a very valuable cross-section of workers and is likely to generalize well beyond university employees.

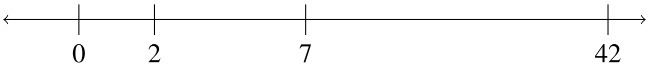

Figure 1. The initial phishing email was sent on day 0. Following phishing emails were send on day 2 (short term), 7 (medium term), and 42 (long term).

We randomly choose 2000 staff members from this list to be subjects in this study. Each subject received all four phishing emails sent between October 24th, 2016 and December 5th, 2016. We limited the study to 2000 staff members for three reasons 1) the tool used to send and track phishing messages (Symantec Blackfin[1]) charges per person, and limiting the number of subjects to 2000 made this study more financially affordable; 2) we were concerned that subjects would talk amongst themselves (as the participants in Kumaraguru et al.'s study did [25]), and by not having all staff members receive messages we somewhat limited the effects of this crosstalk; and 3) preliminary estimates and power analyses suggested that 400 subjects per condition (5 conditions == 2000 total subjects) would be sufficient to detect small-to-medium size effects of the differences in conditions if the expected 30% of subjects clicked on links in the emails.

Phishing Emails

We decided to send out a total of four phishing emails to each subject. After the initial email, we wanted to measure phishing susceptibility at short (2-days), medium (7-days) and long (30-days) timeframes. Due to the unfortunate timing of a security breach that was occupying the time of the IT security and help desk staff, we had to delay the final long-timeframe email to 42 days after the initial email.

To avoid the problem of individuals being repeatedly exposed to the same email (which they might recognize), we chose four different phishing emails to send to each subject. All subjects received the same four emails. However, as Caputo et al. found [5], each email may be different in how difficult it is to identify as a phishing email. If we sent all subjects emails in the same order, which is what Caputo et al. did, then it is difficult to tell if changes in click rates are due to training or if they are due to differences in email difficulty. (Caputo et al.'s second email was much less difficult than the first for their subjects to identify as a phishing email). To address this problem, we randomly assigned subjects to receive the phishing emails in different orders according to a Latin square, with four different orders of emails that are balanced to ensure that for any pair of 2 emails, each email comes before the other in half of the orderings and after the other in the other half of the orderings. The specific Latin square used is in Figure 2.

To select the specific phishing messages that we would send, we began by looking through the list of phishing emails that Cornell University has collected in their PhishBowl website[2].

[1] https://www.symantec.com/products/phishing-readiness
[2] https://it.cornell.edu/phish-bowl

	Email No.			
Group 1	1	2	3	4
Group 2	2	4	1	3
Group 3	3	1	4	2
Group 4	4	3	2	1

Figure 2. Latin square design to counterbalance phishing emails across groups.

We examined these messages and consulted with the IT Services organization at our university, and we identified a set of 12 messages that are very similar to messages that our university has received in the recent past (the last year or so before the study). These 12 phishing emails had a range of different types of messages, including requests from IT services, opportunities for jobs, credit card changes, and extremely vague emails (an email only containing the letters "FYI" linked to a website).

To choose the four emails we used, we conducted an IRB-approved pre-study. We put all 12 emails into a survey and asked participants on Amazon's Mechanical Turk to rate each of the emails on a 5-point Likert scale from “Extremely Fake” to “Extremely Real”. This is a very different population than our subjects; we expected that there may be topics specific to our primary population that the MTurk population might miss, but the general realism of the emails would carry over from one population to the other.

We received 100 valid responses (57% Men, 54% 18-29 years old, paid $1 each for a 5 minute survey) that passed the attention check question. Three of the emails had an average rating above the neutral option. All three emails focused on the same topic: they appeared to be emails from the IT organization alerting users to changes or problems. One was an email about “your mailbox is almost full”, one was a warning about “a new signin attempt from an unusual location”, and the third was a notice about IT “upgrading email accounts”. The email with the next highest average realism also appeared to be from IT: “Click here to revalidate your inbox”. These emails all had similar realism scores, and all appeared to be from the organization's IT department. We choose to use these four emails for the main study. These four, and all 12 of the emails we tested and the results of the pre-study can be seen in our online appendix at `https://osf.io/xxxxxxx` [3].

Finally, we registered four new domain names to use as phishing domains – XXX-it.com, XXX-itservices.com, itservices-XXX.com, it-XXX.com[4]. These domains all appear to be related to our university and its IT organization, but were registered and owned by a third party (the first author) and point to a non-university IP address. Each of the phishing emails was assigned to use one of these domain names. Each email was sent from an email account named ‘IT Services’ with an email address of ‘help@<domain>’. The link in each phishing email was linked to ‘www.<domain>’, which each email being assigned one of the four domains.

Advice and Stories

To test our hypotheses, we need to provide training materials when our subjects click on the link in one of the phishing emails. Specifically, we needed two different types of training materials: 1) a story that teaches people not to fall for phishing, and 2) facts-and-advice that explain phishing and how not to fall for it. However, there is an important potential confound in this design: if the two training materials teach different things (e.g if one of them explains to hover over links to see the URL and the other doesn't), then it is possible that differences between conditions could be because they teach different lessons rather than because of the rhetorical approach (story vs. facts-and-advice).

To address this confound, we conducted a second IRB-approved pre-study to choose the training materials we would use. We first identified three possible stories about phishing incidents that we could use. These stories are real stories from real computer users that came from a database of computer security stories provided by Rader et al. [31]. Second, we identified three different sets of facts-and-advice that explain phishing and provide advice slightly differently. We collected end user training webpages at our university and a set peer universities, and then removed duplicates (many universities copy material from each other, or from the same source, it appears) to arrive at this set of three.

We then went through all six of these texts and brainstormed possible "lessons" that could be learned from these materials, such as "look for https", "type in URLs don't click on them", "phishing is your problem; don't rely on others to protect you", and "misspellings can signal fake emails". We identified a total of 17 lessons that users can learn from these materials. Not all of these lessons are technically accurate or helpful in preventing phishing attacks, but all appeared to us to be present in at least one of the training materials.

We then asked another set of Amazon Mechanical Turk workers to read all six of these training materials, and for each of the 17 lessons, rank whether it is possible to learn that lesson from the training material on a 5-point scale from "Not at all" to "Definitely". We received 51 valid responses that passed the attention check questions. 67% were men, 53% were aged 30-49, and they were paid $2.50 each for a 20 minute survey.

Out of the 17 lessons, we identified 6 lessons (seen in Table 2) that would be directly useful in detecting and responding to the phishing emails we chose in our first pre-study. The other lessons may be useful in general, but do not directly or easily apply to all of the phishing emails we will be sending. For each of the possible lessons, we calculated the average score across the 51 participants' evaluations. The we grouped the lessons into two groups – the 6 lessons that directly affect our phishing emails, and the 11 other phishing lessons – and averaged the scores for each group. From this data, we were able to choose one story and one facts-and-advice to use by

[3] Intentionally left blank for anonymous review. This will be filled in properly for final publication. All materials needed to replicate the study will be made available at that link.

[4] The university's abbreviation has been replaced with ‘XXX’ for anonymous review. We will replace the ‘XXX’ with the actual university acronym for final publication, to reflect the domains actually used.

- Phishing emails can appear to come from your IT department
- Hover over a link to see where it really goes to
- If you accidentally click a link in an email, your identity can be stolen
- Don't click on links in an email
- Phishing is your problem because if you click on the link in a phishing email, it is your information being stolen
- Phishing is when an attacker sends a fake email to you

Table 2. Lessons conveyed by the training materials we selected.

choosing the texts that have the highest average score for the directly useful lessons. Both the story and the facts-and-advice we chose had an average score of 3.7 (out of 5) on the 6 directly relevant lessons, where all four other texts scored below 3.4 for those lessons. There was a difference, however, on the non-relevant lessons; the facts-and-advice contained more lessons than the stories (3.2 vs. 2.5 average score on the 11 non-directly-relevant lessons).

The text of the story and the facts-and-advice used in the main study can be seen in the appendix. Additionally, all six of the texts we tested and the results of the pre-study can be seen in our online appendix at `https://osf.io/xxxxxxx`.

Experts and Non-Experts

For the other manipulation, we wanted to alter the messages to change who appeared to be providing the message: one condition would have an "expert" providing the message, and the other condition would have a "person like me" providing the message.

We began by examining literature about social influence to identify what kinds of people are perceived as "like me". Numerous studies have reported heightened social influence when a person identifies with a source of information and sees that source as "like me". We found that people identify with another person when that other person:

1) is engaged in the same action they are taking [34];
2) is in the same environment (like the same hotel room) [17];
3) have a similar group identify (such as membership in university) [17];
4) believe that others hold similar cultural beliefs [4];
5) are asked to take another person's perspective [10]; or
6) feel empathy for a similar other person over time [7].

After thinking through a number of options, we decided to present our information as coming from "A fellow XXX employee recently clicked on a link in a similar email". This statement makes the message appear to come from someone with similar group membership (3) who recently engaged in the same action as the person who just clicked on a phishing link (1) in a similar situation (2). It may also elicit empathy for the person (6). We believe that this statement will lead the reader to identify with the person providing the anti-phishing message.

To make the message appear to come from an expert, we chose to present the message as coming from "Dr. XXX XXX, a computer security expert"[5]. If subjects choose to search for more information, they would find that he indeed does work for the university and is an expert on computer security. Previous research has suggested that including cues like "Dr." and explicitly labeling him an expert would lead subjects to heuristically evaluate the message as coming from an expert [8, 37, 15].

An advantage of this structure is that we can independently manipulate the expert/non-expert and the story/facts-and-advice. We constructed a training webpage that listed the source of the information on the left side of the webpage (expert or non-expert) and the content of the information on the right side of the webpage (story or facts-and-advice), allowing us to create all four combinations plus a fifth, control webpage that did not include the training material. At the request of the university, we included the university's branding on all training webpages. The final training webpages used in the study are in the online appendix at `https://osf.io/xxxxxxx`.

Ethical Concerns

It is difficult to conduct externally valid phishing studies ethically [14]. One of the biggest challenges is informed consent. If subjects are aware that they are participating in a research study , then they know that they may receive a fake phishing email as part of the study. This can induce two changes in the subjects. First, it can substantially reduce study validity due to changes in behavior. If people know that phishing emails they receive might be part of the study, then they might evaluate these emails differently than they normally would. This is known as "experimenter bias" or the "Hawethorne effect" – people behave differently when they know they are being watched or monitored as part of a study [28]. To avoid this bias, we did not obtain informed consent ahead of the study, following the recommendation of Finn and Jacobson [14]. Instead, we worked to minimize the potential harms of the study by ensuring that subjects did not actually encounter any negative consequences of participating in the study. For example, we did not share subject names or click rates with the university, so there were no possible employment consequences based on subject behavior.

Knowing participation in a phishing research study that involves sending real-looking phishing emails has a second potential change in subject behavior: it is possible that subjects will be more likely to click on phishing emails, believing that the emails are actually benign because they are part of a research study. While our emails actually are benign, subjects have no practical method of distinguishing between our phishing emails and real phishing emails from attackers. This might lead subjects to click on more phishing emails, even though they correctly identified them as phishing messages, because

[5] The name of the first author of the paper was used. Removed for anonymous review.

they believed them to be part of a research study and therefore safe. A colleague at a large corporation reported actually seeing this effect after issuing similar company-sponsored training [16].

While this second effect is another reason to not obtain informed consent ahead of the study, it is even more of a problem for debriefing subjects after the study. If subjects come to believe that phishing messages might be part of the study, then after the study (when *all* phishing messages they receive are real attacks because the study is over) any increase in click rate is a problem. For this reason, we decided to not debrief subjects about their participation in the study. Since the study did not include either informed consent or debriefing, it is possible that subjects never realized that they were part of the study. For this reason, we worked to minimize potential risks to subjects: we pre-tested phishing emails and training messages, we strictly protected subjects identity to avoid any potential negative consequences from participation, we minimized the effort to participate (receiving a total of 4 emails over a 42 day period), and we did not conduct another phishing study or training in the six months after this study. This decision also made it impossible to collect additional data about the subjects, such as gender, age, or employment status, or to survey the subjects as part of the study.

These ethical concerns were discussed carefully with our institution's Institutional Review Board (IRB) and the study was approved by the IRB. Additionally, discussions with our university's IT Services organizations established that we would limit the number of phishing studies conducted on any population.

RESULTS

We conducted our study during the Fall of 2016. Each of the 2000 subjects was randomly assigned to one of the 20 conditions (5 training messages that would appear if they click for each of 4 counterbalanced orders of emails) before the study began. The first email was sent to each subject on Monday, October 24th at around 9:30am. This day of the week and time were chosen as a time that people were likely to be using email, and also because it reflects a common time to receive real phishing emails. The second email was sent on Wednesday, October 26th at 9:30am. The third email was sent on Monday, October 31st at 9:30am, and the final email on Monday, December 5th at 9:30am. All emails were sent at the same time of day to all subjects to minimize any differences due to timing.

Due to a glitch in the query used to select subjects, some people were listed more than once as subjects for the first email. These duplicate entries were removed after the first email, leaving us with a final sample size of 1945 staff members who participated in the study.

Overall, 25.8% of the subjects clicked on at least one of the four phishing emails they were sent (502 people out of 1945). This means that 74.2% (1443) were secure and did not click on a single phishing email that we sent. However, for each individual phishing email, the click rate was much lower. On average, only 8.3% of the emails we sent were clicked on. This means that for a single phishing email sent to a single person, there is only a 8.3% chance that they will click on it (assuming the phishing email is similarly difficult to the ones we sent). This number is lower than reported in prior field studies (Table 1) [25, 5, 12].

Email	Percent Clicked
Mailbox is almost full	6.0%
New Sign-in Attempt	7.0%
Upgrading Email accounts	12.1%
Re-validate your mailbox	7.7%

Table 3. The subject lines of the four different phishing emails that we sent, and the percentage of subjects who clicked on each one. Every subject received each email exactly once.

Day	# Clicked	Percent Clicked	Repeat Clicks
0	228	11.7%	0%
2	143	7.35%	21.7%
7	125	6.43%	35.2%
42	146	7.51%	44.5%

Table 4. Fewer people clicked on phishing emails in later days, though there was a slight increase after a month. The last column indicates, out of the people who clicked, what percentage of them had previously clicked on the link in one of our phishing emails (and therefore previously received training).

Some Emails are More Difficult

We sent four different phishing emails. Everyone received each email, though our latin square design had people receiving them in different orders. This allows us to tell which email is the most difficult.

We found that one email – about upgrading email accounts – was significantly more difficult than the other three (Table 3). One possible explanation is that the university was working on upgrading email accounts at the time, and it is possible that subjects had heard about this effort. The phishing email we sent, though, was not from as official university email address and was not part of this effort. This suggests that the latin square design was very important as an experimental control.

Overall, The Training Worked

Each person received four phishing emails, on days 0, 2, 7, and 42. For the first week, subjects collectively got better and clicked on fewer emails with each subsequent phishing mail. However, after a month passed, they regressed slightly. Table 4 shows this pattern.

Notice that the majority of clicks were not from subjects who had previously clicked one of our phishing links; instead, most people who clicked on a link in a given day were new victims.

Few People Clicked More than Once

Caputo et al. [5] found that a large number of people are "always clickers" – they click on all phishing emails sent. Caputo et al expressed concern that these always clickers might not respond to training at all. While over 10% of their subjects always clicked, in our study we found only 6 people out of 1945 that were always clickers – less than 1%.

Type of Person	Subjects	Percent	Caputo et al. [5]
Never Clicked	1443	74.19%	21.7%
Clicked Only Once	396	20.40%	38.7%
Clicked After Training	106	5.56%	39.5%
Always Clicked	6	0.31%	10.7%

Table 5. Percent of subjects who never clicked or always clicked.

Very few subjects clicked on the link in more than one of our phishing emails. Since subjects only received training messages after clicking on one of our phishing links, any subject who clicked more than once clicked after receiving the training. Only 5.45% of the subjects (106 people) clicked after receiving training.

We found a very large group of "never clickers" – 1443 people. In our sample, this group is much larger than in prior research [5]. It is good to see this group growing over time.

Training Effectiveness

The basic effectiveness of training for the organization is evidence above: the overall percentage of clicks declined over time, and was much lower after one week (about 45% lower) than at the beginning of the study.

Organizational effectiveness is different from and distinct from individual improvement. The organization as a whole can improve because people talk about the phishing emails they receive and the training they receive, which can lead others to learn from those stories and be trained indirectly [31, 25].

This indirect training effect was given as a potential confound in one previous study [25]. To avoid this confound, we only included in the study a randomly selected 2000 out of the approximately 7500 eligible staff members, hoping that this would limit the spread of information. Additionally, we focus our analysis specifically on the people who received training, and their subsequent responses to phishing messages.

Who was Trained?

502 people received training as part of this study, which is 25.8% of the people who were selected to be in the study. They only received training if and when they clicked on one of the phishing emails. Each person only received training once, even if they clicked on more than one phishing message. This was done to prevent them from learning that phishing messages are generally safe to click on because they come from the university. If they clicked a second time, they received a complete blank webpage.

Of those people, we are able to measure the effectiveness of the training for everyone who clicked on one of the first three emails (421 people). This is because people who first clicked on the fourth email and received training did not have another opportunity to use their training as part of this study.

For this group of people of 421 people, 25.2% clicked on one of the later emails (106 people). This is a noticeably higher percentage of clicks than the overall click rate. This is not surprising; by clicking on a phishing email, they have revealed themselves to be susceptible to phishing attacks and thus more likely to click on a phishing email than the average person.

	Advice	Story
Expert	18%	34%
Person Like Me	25%	24%
Control	23%	

Table 6. The percentage of people who clicked on a phishing email *after* receiving a training message. Subjects only received training after initially clicking.

	Clicked	Clicked Again	Percent
Advice from an Expert	76	14	18%
Control	81	19	23%
Story from a Person Like Me	89	21	24%
Advice from a Person Like Me	87	22	25%
Story from an Expert	88	30	34%

Table 7. Number of People who ever clicked after receiving training. Percentages are the same as in Table 6. $\chi^2(4, N = 421) = 6$, $p = 0.22$

Types of Training

Tables 6 and 7 compare the results of the different types of training conducted. These results suggest that the traditional training – providing direct advice about how to deal with phishing emails, and having it originate from people who are seen as experts – is the best strategy for training people. It led to a 21% decrease in clicks compared to the Control condition. However, there is an important caveat here: if the exact same advice is seen as coming from a peer ("A fellow university employee"), then the advice no longer works, and leads to a very similar click rate as not training subjects (our control condition).

Stories, however, seem to have the opposite pattern. If a story about a phishing incident is told by an expert, then that actually increases the likelihood of clicking on a phishing email link. On the other hand, if a story is told by someone similar to them, it doesn't not seem to have much effect; it doesn't increase or decrease clicking relative to the control group.

These results indicate an interaction effect, as can be seen in Figure 3. Stories are more effective when subjects perceive them as coming from people similar to them, whereas facts-and-advice are more effective when coming from experts.

DISCUSSION

Our major finding is the interaction effect illustrated in Figure 3: that facts-and-advice training leads to lower likelihood of clicking on a phishing link when appearing to from an expert than from a peer; and a story that conveys the same security lessons leads to lower likelihood of clicking in the opposite situation. We have identified two potential explanations for this pattern in our data, but this current study cannot distinguish between these two explanations.

First, this could be a moderation effect [1]. The source of the information could moderate how people interpret the information; stories are taken more seriously from peers, and

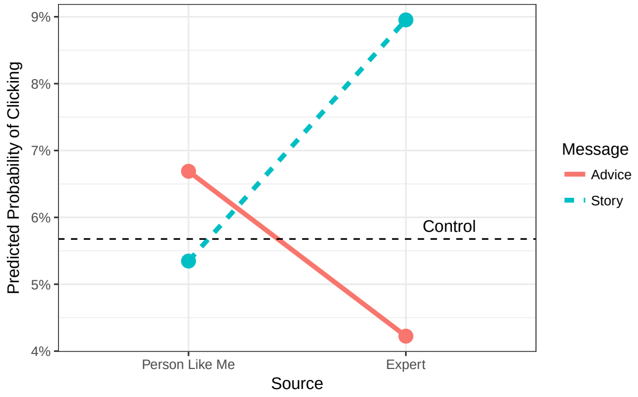

Figure 3. Estimated Likelihood of clicking on a phishing link after receiving different types of training. Estimates are predicted values from a multi-level logistic regression model that controls for elapsed time and individual differences. Regression model output can be found in supplemental materials.

advice is absorbed better from experts. Moderation effects are common in persuasion (e.g. [29]), and source of information is frequently an important moderator [29]. Specifically in the context of computer security and phishing training, the source of the information has the potential to be a very important moderator for whether and how the training ends up affecting user behavior.

Second, our data is also consistent with an alternative explanation: that experts have a strong ability to influence behavior, both positively and negatively, but peers have little influence. Facts-and-advice from experts showed the lowest post-training click rate, but stories from experts produced a post-training click rate much higher even than our control group that didn't receive training. Both the facts-and-advice and the story manipulations contained the same security lessons, but it is possible that the lessons in the stories backfired when coming from experts. This could be similar, for example, to a well-documented case in the US National Petrified Forest where signs discouraging people from stealing actually lead to increases in theft because they made it seem like everyone was doing it [6].

On the other hand, both conditions where content appeared to come from people-like-me had post-training click rates with 1–2% of the control condition. It is possible that training coming from peers is simply ignored, and has little effect relative to the large effects present for expert-presented content. Even when providing content that contains the same lessons, experts have a large influence can cause either an increase or a decrease in click rates on phishing, but peers have less influence.

While both explanations are interesting, our data cannot conclusively distinguish between these two explanations. We only have two data points – one facts-and-advice, and one story – that we used to compare these two sources. There is a much larger possible space of training materials, and these two data points are not enough to distinguish between our alternative explanations. Both explanations are consistent with past research on persuasion [6, 29], but lead to different conclusions. Future research will be needed to tease apart these explanations.

A Statistical Challenge for Embedded Training Research

This study also revealed a particular statistical challenge that is common to other similar studies that use embedded training [26] to improve phishing. Our study had 1,945 subjects – larger than most other similar studies, as can be seen in Table 1. However, because we used the industry standard embedded training, only subjects who clicked on a phishing link in one of our first 3 emails actually received training in time for us to measure its effect. This reduced our effective N from 2,000 to 421, or about 20% of our original subject pool.

As society's awareness and training about phishing increases, we are seeing a decline in people's willingness to click on phishing emails. Our average click rate of 8.3% per email (11.7% if you only consider the first phishing email) is lower than past studies, and reflects a societal decline in phishing susceptibility over time (Table 1). This is making it much more difficult to use inferential statistics to study phishing training because this lower rate affects the number of data points for the statistics.

In our study, one condition had almost twice as many clicks as another condition (34% vs 18%, Table 7), but a χ^2 test of the different click rates between conditions was not statistically significant. This is a large practical effect – training that reduces clicks by 50% is extremely valuable – but is not statistically significant. Past studies have also had this problem, such as the study by Caputo et al. [5]. It is invalid to conclude that a statistically non-significant result is actually no effect; we still believe that there is an effect and our results represent our best estimate of that effect. However, we cannot necessarily rule out that these differences are due to chance. And this problem will only get worse as our training methods and society's awareness of phishing improves.

CONCLUSION

Phishing remains one of the more widely exploited human vulnerabilities today. Training users to recognize and avoid clicking on links in phishing emails is a large and important business today. We compared two major methods of conducting this training: providing facts-and-advice about phishing, or providing stories of previous victims of phishing. While both can contain the same lessons for end users, we found a surprising interaction effect: facts-and-advice led to lower click rates when appearing to come from an expert, but stories led to lower click rates when appearing to come from peers rather than experts. We discussed competing potential explanations for this interaction effect, but cannot concretely explain it.

ACKNOWLEDGEMENTS

APPENDIX

TRAINING MATERIALS

Facts and Advice

Phishing is an online scam involving email messages appearing to be from a trusted source. A type of phishing, called spear phishing, is especially problematic.

Spear phishing is a technique that con artists use to specifically target individuals or companies and gain access to private information or accounts.

With spear phishing, hackers disguise themselves as a trusted source by sending an email with a request to provide personal information, such as log in and password information. When the person gives the information by replying to the email or via a website link provided, the criminal goes into the account and takes what they want.

Watch for:

1. *The email urges you to take immediate action.* Often, a phishing email tries to trick you into clicking a link by claiming that your account has been closed or put on hold, or that there's been fraudulent activity requiring your immediate attention. To be safe, log into the account in question directly by visiting the appropriate website, then check your account status.
2. *The hyperlinked URL is different from the one shown.* The hypertext link in a phishing email may include the name of a legitimate bank. But when you hover the mouse over the link (without clicking it), you may discover in a small pop-up window that the actual URL differs from the one displayed and doesn't contain the university's name.
3. *Be wary of messages demanding immediate response and requesting passwords, bank accounts, or threatening to suspend or terminate your account.* Look at the sender's email address. Does it make sense? Is it from someone you know? If you don't know the person or the email account is not associated with the actual organization, look up the number for the institution and contact them to verify its authenticity. Do not use any phone numbers provided by the suspected sender.

Phishers could take stolen account credentials and sell them to criminals who could use your email or account to send huge volumes of spam. They could also gain access to your personal and financial information.

Story

"Sometimes frauds will target university email addresses to trick them into giving up information about themselves. I made the mistake of offering up information even after hearing this. I got a message from the "IT department" requesting that I verify my account information, otherwise my account will be suspended.

"Stupid me, I should have known that it was a trick. When I clicked on the email, it took me to a website that wasn't really my university. I had to end up canceling my account and getting a new one, changing my password, etc. It was pretty embarrassing.

"I quickly wizened up and have since never ever been a victim again. Now I hover over links to see where they link to. I won't be fooled twice."

REFERENCES

1. Reuben Baron and David Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51, 6 (December 1986), 1173–1182.
2. Roy F Baumeister, Liqing Zhang, and Kathleen D Vohs. 2004. Gossip as Cultural Learning. *Review of General Psychology* 8, 2 (2004), 111–121.
3. Zinaida Benenson, Freya Gassmann, and Robert Landwirth. 2017. Unpacking Spear Phishing Susceptibility. In *Targetted Attacks (Financial Cryptography and Data Security Workshops)*, Markus Jakobsson (Ed.). Springer.
4. Cristina Bicchieri, Jan W Lindemans, and Ting Jiang. 2014. A structured approach to a diagnostic of collective practices. *Frontiers in Psychology* 5 (2014), 93.
5. Deanna D Caputo, Shari Lawrence Pfleeger, Joshua D Freeman, and M Eric Johnson. 2014. Going Spear Phishing: Exploring Embedded Training and Awareness. *Security & Privacy, IEEE* 12, 1 (Jan. 2014), 28–38.
6. Robert B Cialdini, Linda J Demaine, Brad J Sagarin, Daniel W Barrett, Kelton Rhoads, and Patricia L Winter. 2006. Managing social norms for persuasive impact. *Social Influence* 1, 1 (March 2006), 3–15.
7. Jonathan Cohen. 2001. Defining Identification: A Theoretical Look at the Identification of Audiences With Media Characters. *Mass Communication and Society* 4, 3 (Aug. 2001), 245–264.
8. Richard Crisci and Howard Kassinove. 1973. Effect of Perceived Expertise, Strength of Advice, and Environmental Setting on Parental Compliance. *The Journal of Social Psychology* 89, 2 (Apr 1973), 245–250. DOI:http://dx.doi.org/10.1080/00224545.1973.9922597
9. Sauvik Das, Tiffany Hyun-Jin Kim, Laura A. Dabbish, and Jason I. Hong. Submitted. The Effect of Social Influence on Security Sensitivity. USENIX Association, 143–157. https://www.usenix.org/conference/soups2014/proceedings/presentation/das
10. Anneke de Graaf, Hans Hoeken, José Sanders, and Johannes W J Beentjes. 2012. Identification as a

Mechanism of Narrative Persuasion. *Communication Research* 39, 6 (Dec. 2012), 802–823.

11. Rachna Dhamija, J D Tygar, and Marti Hearst. 2006. Why phishing works. In *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Request Permissions, New York, New York, USA, 581–590.

12. Aaron Ferguson. 2005. Fostering E-Mail Security Awareness: The West Point Carronade. *EDUCAUSE Quarterly* (2005).

13. Bryan Ferrario. 2015. The True Cost of Phishing Attacks. Alliance Web Security Blog. (November 2015). https://www.alliancewebsecurity.com/true-cost-phishing-attacks/

14. Peter Finn and Markus Jakobsson. 2007. Designing Ethical Phishing Experiments. *IEEE Technology and Society Magazine* 26, 1 (2007).

15. B. J. Fogg, Preeti Swani, Marissa Treinen, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, and et al. 2001. What makes Web sites credible? *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '01* (2001). DOI: http://dx.doi.org/10.1145/365024.365037

16. Viabhav Garg. 2015. Personal Communication. (2015).

17. Noah J Goldstein, Robert B Cialdini, and Vladas Griskevicius. 2008. A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels. *Journal of Consumer Research* 35, 3 (Oct. 2008), 472–482.

18. Rick C. Hodgin. 2007. Phishing Cost the U.S. $3.2 Billion in 2007. Tom's Hardware. (December 2007). http://www.tomshardware.com/news/phishing-cost-u-s-3-2-billion-2007,4576.html

19. Tom N Jagatic, Nathaniel A Johnson, Markus Jakobsson, and Filippo Menczer. 2007. Social phishing. *Commun. ACM* 50, 10 (Oct. 2007), 94–100.

20. Rob Joyce. 2016. Disrupting Nation State Hackers. In *USENIX Enigma*. San Fransisco, CA.

21. Gary A Klein. 1998. *Sources of Power*. MIT Press.

22. Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. 2009. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the Symposium on Usable Privacy and Security*. ACM, New York, New York, USA, 3.

23. Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007a. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the ACM Conference on Human Factors in Computing (CHI)*. ACM, New York, New York, USA, 905–914.

24. Ponnurangam Kumaraguru, Yong Rhee, Steve Sheng, Sharique Hasan, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. 2007b. Getting users to pay attention to anti-phishing education: evaluation of retention and transfer. In *Proceedings of the anti-phishing working group eCrime Researchers Summit*. ACM, New York, New York, USA, 70–81.

25. Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. 2008. Lessons from a real world evaluation of anti-phishing training. In *eCrime Researchers Summit*. IEEE, 1–12.

26. Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. 2010. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)* 10, 2 (May 2010), 7–31.

27. Jess Nelson. 2016. Email Phishing Attacks Estimated to Cost $1.6M Per Incident. Email Marketing Daily. (January 2016). https://www.mediapost.com/publications/article/267680/email-phishing-attacks-estimated-to-cost-16m-per.html

28. H. M. Parsons. 1974. What Happened at Hawthorne?: New evidence suggests the Hawthorne effect resulted from operant reinforcement contingencies. *Science* 183, 4128 (Mar 1974), 922–932. DOI: http://dx.doi.org/10.1126/science.183.4128.922

29. Richard E Petty and John T Cacioppo. 1986. *Communication and persuasion : central and peripheral routes to attitude change*. Springer.

30. Emilee Rader, Alcides Velasquez, Kayla D Hales, and Helen Kwok. 2012a. The gap between producer intentions and consumer behavior in social media. In *GROUP '12: Proceedings of the 17th ACM international conference on Supporting group work*. ACM Request Permissions.

31. Emilee Rader, Rick Wash, and Brandon Brooks. 2012b. Stories as informal lessons about security. In *SOUPS '12: Proceedings of the Eighth Symposium on Usable Privacy and Security*. ACM, New York, New York, USA, 1.

32. Emilee J Rader and Rick Wash. 2015. Identifying patterns in informal sources of security information. *Journal of Cybersecurity* (2015), tyv008.

33. Zulfikar Ramzan. 2010. Phishing Attacks and Countermeasures. In *Handbook of Information and Communication Security*, Peter Stavroulakis and Mark Stamp (Eds.). Springer, Chapter 23.

34. Raymond R Reno, Robert B Cialdini, and Carl a Kallgren. 1993. The transsituational influence of social norms. *Journal of personality and social psychology* 64, 1 (1993), 104–112.

35. Stuart E Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. 2007. The Emperor's New Security Indicators. In *SP '07: Proceedings of the 2007 IEEE Symposium on Security and Privacy*. IEEE Computer Society, 51–65.

36. Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd Symposium on Usable Privacy and Security (SOUPS)*. ACM Press, New York, New York, USA, 88.

37. Shyam Sundar. 2008. “The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. In *Digital Media, Youth, and Credibility*, Miriam J. Metzger and Andrew J. Flanagin (Eds.). MIT Press, 73–100.

38. Rick Wash. 2010. Folk models of home computer security. In *SOUPS ’10: Proceedings of the Sixth Symposium on Usable Privacy and Security*. ACM, New York, New York, USA, 1.

39. Min Wu, Robert C Miller, and Simson L Garfinkel. 2006. *Do security toolbars actually prevent phishing attacks?* ACM, New York, New York, USA.