# Measuring the Effectiveness of Embedded Phishing Exercises

*Hossein Siadati* † *hossein@nyu.edu*
*Sean Palka* ‡ *palka_sean@bah.com*
*Avi Siegel* ‡
*Damon McCoy* † *mccoy@nyu.edu*

† *New York University*
‡ *Booz Allen Hamilton*

## Abstract

Embedded phishing exercises, which send test phishing emails, are utilized by organizations to reduce the susceptibility of its employees to this type of attack. Research studies seeking to evaluate the effectiveness of these exercises have generally been limited by small sample sizes. These studies have not been able to measure possible factors that might bias results. As a result, companies have had to create their own design and evaluation methods, with no framework to guide their efforts. Lacking such guidelines, it can often be difficult to determine whether these types of exercises are truly effective, and if reported results are statistically reliable.

In this paper, we conduct a systematic analysis of data from a large real world embedded phishing exercise that involved 19,180 participants from a single organization, and utilized 115,080 test phishing emails. The first part of our study focuses on developing methodologies to correct some sources of bias, enabling sounder evaluations of the efficacy of embedded phishing exercises and training. We then use these methods to perform an analysis of the effectiveness of this embedded phishing exercise, and through our analysis, identify how the design of these exercises might be improved.

## 1 Introduction

Phishing is a threat that has been an initial attack vector in several recent data breaches [10, 22]. Thus, organizations deploy filters and also human training defense strategies, such as planned phishing awareness training and unannounced embedded phishing exercises. Previous research has found that unannounced embedded phishing exercises, in which test phishing emails are sent to employees to see if they will act on them, can provide a "teachable moment." After a user falls for a test phishing email, he or she will be more receptive to training [14].

While most, but not all, prior research agrees that embedded phishing exercises followed by awareness training reduce employee's susceptibility to phishing [2, 13, 14], correctly designing and soundly evaluating the effectiveness of these exercises is not easy. Little or no research has explored possible factors that might bias results, such as differing levels of persuasiveness of the phishing message or other issues. Without guidance based on systematic studies and frameworks, our observation is that companies have had to create their own ad-hoc evaluation methods for their phishing exercises.

In this work, we conduct an analysis of data from a real world embedded phishing exercise that tested 19,180 participants from a single organization, and utilized 115,080 test phishing emails distributed in six rounds over eight months. In this dataset, we observed test phishing emails with average raw click-through rates that varied from 0% to 40%. This variance in the pervasiveness of phishing emails indicates that an evaluation metric using raw click-through rates for each round, would produce an inaccurate effectiveness measure.

The first part of our study focuses on developing methodologies for correcting possible sources of bias, which enables us to make a more sound evaluation of the effectiveness of embedded phishing exercises. A reliable objective technique for measuring these biases, and for determining the best techniques to mitigate their influence are necessary for drawing meaningful conclusions about how to evaluate and design phishing exercises.

Our analysis shows that training can have a significant effect on decreasing the susceptibility of the users to phishing schemes. In the dataset that we studied, on average, training decreased the phishing click-through rate by 40%. However, this decrease was only observed when using more persuasive phishing emails. We found that embedded training of the type studied in this work is likely not useful in providing protection for vulnerable users who are easily deceived by unpersuasive phishing emails. Our improved evaluation metric suggests that current designs of the company's embedded phishing exercises are suboptimal with respect to maximizing the effectiveness of training.

## 2 Background

In this section, we describe the common design of large-scale embedded phishing exercises performed by industry. To provide some context to our study, we then examine related work on training methods and phishing exercises from academic literature.

### 2.1 Designing Phishing Exercises

We describe components and process of embedded phishing exercises based on interviews with three medium-sized US- based companies that regularly conduct these exercises. As part of this description, we illuminate some of the less understood aspects of designing these exercises. While we found that this design was relatively consistent across all three of the interviewed companies [1], we make no claims that the design is fully representative of how all embedded phishing exercise are conducted.

**Content creation.** A limited set of phishing emails are manually created by subject matter experts for each embedded phishing exercise. These phishing emails have different themes, such as receiving a fax or resetting a password, to entice users to click on the links. A single phishing email will be sent to a number of people, and may later be reused across rounds to save the cost of content generation [2].

**Grouping.** In the three companies we studied, two partitioned employees into cohorts at random, by department, or job function, while one did not create partitions. All members of a cohort were sent the same test phishing email in a round, while members of other cohorts could received a different email. In the third company we studied, all employees were sent the same test phishing emails.

**Phishing awareness training.** When an employee clicks a link in a test phishing email, he is immediately redirected to a website that informs him that he has fallen for a phishing email. The employee is then linked to a phishing awareness training program, thus turning an error into a teachable moment.

**Multiple rounds.** All of the embedded phishing exercises we examined were conducted in multiple rounds in order to assess the effectiveness of this training at reducing both an individual's or an organization's susceptibility to phishing attacks. These rounds were spaced out over time, with a few weeks to a few months between rounds.

**Evaluation.** Soundly assessing the effectiveness of these embedded phishing exercises is a challenging problem that is done differently by each organization.The effectiveness of the phishing exercises is primarily evaluated on changes in the click-through rate of first and last rounds of exercise. A campaign is considered effective if the average-click-through rate is significantly reduced. However, because of the difference in the persuasiveness of phishing emails (measured based on click-through rate of average untrained users) given in different rounds of the exercise, using the average raw click-through rate can be misleading. In this paper, we uncover the biases of current measurements, and provide an appropriate process to account for them.

### 2.2 Related Work

#### 2.2.1 User-centered anti-phishing methods

Users should be trained to safely handle a phishing email that bypasses the automatic detections. Combating phishing attacks through education and training was initially proposed by Liao and Luo [16]. Several approaches to phishing training have been proposed [7, 9, 15, 18, 20, 23]. Embedded phishing exercise attempts to train users by sending test phishing emails[3] to employees, typically without any prior notification to better simulate the conditions of an actual phishing attack. Unannounced exercises that include a training component have been shown to be more effective than other methods likely because this method provides phishing training at a *teachable* moment, right after the users click on a phishing link [12]. Embedded phishing exercise is well received by companies for training as well as measuring the resiliency of their employees against these attacks.

Previous works are limited in measuring the effect of phishing training because of the small sample size, as well as the number of rounds in the exercise, and the types of phishing emails used in experiments. In comparison, this work studies the effect of embedded training on close to 20,000 employees of an organization in multiple rounds with various phishing email types.

#### 2.2.2 Methods of evaluation

The general structure for an evaluation method is to surround a training phase with two tests, one before and the other one after the training. A change in the percentage of users who click on phishing emails before and after training is used to judge if vigilance against phishing emails has changed. Where approaches differ in how similar the before and after tests should be, the common sense is that training should have a similar level of difficulty, or some sort of normalization should be used to make the results comparable [20]. In other

[1] Only one of these companies provided us with data from their phishing exercise.

[2] These emails are whitelisted to reduce the likelihood of reused test phishing emails being filtered.

[3] The content is very similar to phishing email but it is harmless.

words, an evaluation must consider how persuasive phishing emails are to the subject before and after training. However, many studies have not considered this factor in their evaluations [7, 9, 19]. Others have used different strategies to cancel potential biases:

**Replicated tests.** In this method, the same set of phishing and benign emails are given before and after the training [21].

**Swapping halves.** In this method, the test set is split into two subsets, A and B, with the same number of phishing and benign tests, that are given in different order to subjects of different groups [20].

**Counterbalance schedule.** This method divides participants randomly and equally into groups, and the tests are scheduled in a way that each phishing email is given to one partition in each round. For any given day of the study, each of the email types are sent to an equal number of participants [11].

**Pre-selection.** In this method, phishing emails with equal level of persuasiveness are selected for evaluation of user performance. Persuasiveness of phishing tests are gauged by a separate pool of users, during the design process. Real pre-tests and iterations required for that add complications of a campaign design, and therefore, pre-selection is rarely used in practice. However, surveying experts, instead of a real test, is used for pre-selection, but is shown to be unreliable [4].

The phishing training campaign we studied did not use any of the methods described above. Instead, phishing emails of various topics (e.g., celebrity, shipping) were assigned randomly to different groups. This creates difficulty for fair evaluation of performance of employees. In this paper, we propose normalization methods suitable for analysis of such flexible scheduling and assignment of phishing emails in the training exercises. Normalization, or weighting of responses, has often been used in survey- based studies involving phishing [3, 5, 17]. Ratio-based normalization, similar to the one employed by Hage et. al to evaluate the effectiveness of academic courses and standardized tests [6, 8] are frequently used in practice.

# 3 Campaign Design and Data

This study is based on data acquired from a large-scale phishing training campaign conducted in a medium-size company. The structure of this campaign is by no means optimal and reflects the limited resources made available by companies for conducting this type of training. It also, to some extent, highlights the current lack of best practices for the design of large-scale phishing training campaigns, as a result of the limited resources allotted for administering such a campaign. Some key facts about the phishing training campaign are as follows:

- It Included 19,180 participants
- Participants were partitioned into 32 different groups
- A total of 28 different test phishing emails with varying levels of persuasion were used
- The training campaign was conducted in six different rounds, spread out over eight months

After this paper is accepted for publication, we will publicly release a copy of this raw anonymized data to enable further exploration of the data by other researchers.

## 3.1 Phishing Training Campaign Structure

**Population.** Participants were selected from an initial list of 23,062 email addresses, pruned down to 19,180 individual addresses after removing ones that were either invalid, general (e.g., distribution lists), etc. The participants were divided into 32 different test groups, based on recommendations from the organization to minimize impact on business operations. The average size of each group was 599, though they varied between 68 to 937 employees per group.

**Phishing Emails.** The emails were developed manually by phishing training experts to represent common types of phishing messages that had been detected in the wild. A total of 28 different phishing emails were designed, ranging from reward promises (e.g., free cruise) to bogus receipts for online transactions.

**Schedules.** Each group of employees received 6 different test phishing emails over the course of 8 months at almost equal intervals. The assignment of the type of email and their order was randomized. We refer to rounds by a letter between A to F.

**Training.** Immediately after clicking on a phishing link, the subject was offered training. They were redirected to an training page, including (I) a few slides on how to identifying suspicious elements in a phishing email, and (II) an interactive simulated email client (Inbox) with various emails customized to that user. They had to identify a certain number of suspicious elements to complete the training.

**Ethical considerations.** The phishing training campaign was performed at the request of company management with a goal to improve the organization's resilience to phishing attacks. All phases of the campaign were done with the permission of, and in coordination with, the system owners. The protocol of receiving and analysis of the data was also approved by the researcher's institutional review board (IRB).

## 3.2 Data Collected

Our research is based on anonymized data provided by the group that conducted the phishing campaign. We ex-

| Email | Exercise Round | | | | | |
|---|---|---|---|---|---|---|
| | A | B | C | D | E | F |
| Celebrity | 0 | n/a | 0.1 | n/a | 0.1 | n/a |
| Sports | 0.3 | n/a | 0 | 0.1 | 0 | n/a |
| Newsletter | n/a | n/a | 0.2 | n/a | 0.4 | 0.1 |
| Dragon | 0.7 | 0.3 | 0.4 | n/a | 0.1 | n/a |
| Big Box | n/a | n/a | n/a | 0.8 | n/a | n/a |
| Bank 2 | 0.8 | 0.2 | 0.5 | n/a | n/a | n/a |
| Bank 1 | n/a | n/a | n/a | 0.8 | 2.4 | n/a |
| Warehouse | n/a | 2.5 | 2.6 | n/a | n/a | 0.1 |
| Certify | 3 | 1.6 | n/a | n/a | 0.7 | 0.8 |
| NACA | 3.6 | 3.1 | 2.5 | n/a | 1.8 | n/a |
| Charity | n/a | n/a | n/a | 2.8 | 1.6 | 0.7 |
| Malware | 3.5 | 2.9 | n/a | n/a | n/a | n/a |
| Outfitters | n/a | n/a | n/a | 4 | 3.2 | 3.2 |
| Federal | n/a | 5.7 | n/a | 4 | n/a | n/a |
| Secure Mail | n/a | 6.3 | n/a | 2.5 | n/a | 4.9 |
| AGCE | n/a | 7.5 | n/a | n/a | 2.8 | 5.5 |
| Bazaar | 6.9 | n/a | 3.8 | n/a | 2.2 | n/a |
| Cellular | n/a | 9 | n/a | n/a | 2.7 | 1.1 |
| IQ Test | n/a | 6.7 | 6.4 | n/a | 10.4 | n/a |
| Tax | 9 | n/a | n/a | n/a | n/a | 3.2 |
| Password | 10 | 7.3 | n/a | n/a | 2.8 | n/a |
| Funds | 9.9 | 4.9 | 5.1 | n/a | n/a | n/a |
| Domain | 5.3 | n/a | 7.8 | n/a | n/a | n/a |
| Fax 2 | n/a | 26.3 | 22.5 | 19.4 | n/a | 15.6 |
| Shipping | n/a | n/a | n/a | 21.2 | 13.2 | 9 |
| Fax 1 | n/a | 27.1 | n/a | 17.2 | n/a | 1.6 |
| Complaints | 26.3 | n/a | 17.9 | 17.8 | 12.8 | 3.5 |
| Order | 28.5 | 36.3 | n/a | n/a | n/a | 10.7 |
| Overall Clicks | 1507 | 1483 | 1035 | 1114 | 715 | 983 |
| Overall New Clicks | 1507 | 1362 | 794 | 786 | 478 | 705 |
| Overall Click rate | 7.9 | 7.7 | 5.4 | 5.8 | 3.7 | 5.13 |

**Table 1:** Combined click-through rate for each phishing email across all populations for six rounds of exercises. Some fields are marked as "n/a". This indicates that the emails were not given to any of the groups in that round. Last rows summaries the number of clicks on phishing emails as well as overall click-through rate in each round.

plain here the type of data collected and offer aggregated views of the results.

**Phishing Clicks.** This exercise focused only on the susceptibility of users to clicking on links within test phishing emails.

This is standard in phishing measurement since previous works have shown that most of the users who click on phishing links will reveal their credentials [12].

Therefore, we refer to a user who clicks on a link as *phished*, and those who do not as *not-phished*.

Table 1 shows the combined click-through values for each type of email used in each exercise. Alternatively, Figure 1 shows the range of click-through rates that different emails receive thorough the phishing campaign.

**Training.** The phishing exercises were designed around training users in "teachable moments," that occur when a mistake—in this case a click on a link— is made. The dataset we were provided included the behavior of each

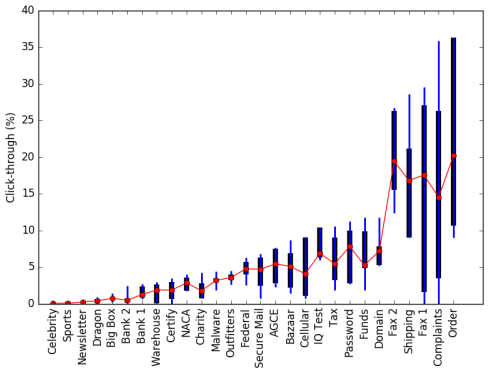


**Figure 1:** Click-through value ranges for the 28 phishing emails used in the study. The thiner bar shows the range of click-through for individual groups. The thicker bar shows the range for click-through rate of rounds. Red dots show the weighted average click-through over the rounds. Emails are sorted based on the maximum click-through rates.

participant with respect to training. Specifically, it indicates if a users was trained, meaning they were offered training and completed it, as well as how long it took for them to finish the training. We label a phished subject that was immediately notified and took the training as *Trained*. Users who visited the training page but did not take the training are considered as *Notified*. Due to settings on the browser or network where users clicked on the phishing link, some could not reach the phishing page[4]. We label such subjects *Not-notified*. Table 2 lists the percentage of each category in different rounds.

| Round | Trained | Notified | Not-notified |
|---|---|---|---|
| A | 28.7 | 6.1 | 65.2 |
| B | 26.3 | 6.2 | 67.5 |
| C | 72.1 | 7.7 | 20.2 |
| D | 76.0 | 5.6 | 18.2 |
| E | 71.1 | 7.4 | 21.5 |
| F | 71.0 | 7.9 | 21.1 |

**Table 2:** Percentage of trained, notified, and not-notified subjects in different rounds relative to the total number of phished subjects in that round.

## 3.3 Evaluation Challenges

The major questions regarding the effectiveness of a phishing exercise include: "Did the organization and/or individuals within it become more resilient to phishing attacks? And, if so, by how much?"

Raw data on the performance of individuals during these exercises can not be directly used to answer these questions with any real certainty. The following example from the dataset explored in this paper illustrates

[4]The phishing URL redirected them to a training page where they were notified of phishing and offered training

the challenge. In this dataset, group 20 started out with 35.9% of participants being phished in exercise A. The phished rate then dropped to 1.9% and 2.4%, respectively, in exercise B and C, before increasing to 19.4% in exercise D. Similar fluctuations in performance are observed for most of the other groups. The reason for these irregularities become evident when we consider the click-through rate of each type of email throughout the six phishing exercises. For example, when comparing the "Sports" and "Complaints" phishing emails in Table 1, we see that the former never receives more than a 0.3% click-through rate, while the latter achieves a click-through rate of up to 26%. This shows that the changes in click-through rates in different rounds are correlated with the inherent persuasiveness of the phishing emails.

Therefore, drawing conclusions about the effect of exercises based only on trends in the raw click-through rates of different emails can lead to erroneous conclusions.

This observation suggests that the click-through rate should be normalized based on the persuasiveness of the content of phishing email text to produce a sounder analysis. The catch is that normalizing the click-through rate requires knowing the persuasiveness of the phishing emails. Lacking an a priori measure of email persuasiveness, we must devise a metric to estimate that based on the given data. In the next section, we will develop methods for estimating phishing email persuasiveness and click-through rate normalizations to more soundly evaluate the results of phishing training exercises.

## 4 Email Persuasiveness

As part of our methodology for this study, we needed to find a way to quantify the *persuasiveness* of phishing emails for average untrained users. In this work, higher click-through rate on phishing links in a phishing email means higher persuasiveness of the phishing email. In the phishing training campaign studied, phishing emails were given in different rounds. Therefore, an email's persuasiveness can not be computed simply by averaging over the click-through rates of all data points related to it. Such an evaluation is sensitive to the round in which the emails are given. Considering the maximum click-through rate throughout phishing training is also not appropriate since this measure is overly sensitive to the performance of a group.

Our approach to compute the persuasiveness of a phishing email uses regression over performance of different groups in different rounds in response to the given phishing email. In the absence of prior analysis on trend of changes of click-through rates, we assume that it is linear and therefore use a simple linear regression over click-through rates. Developing more accurate models is an area for future exploration. For a given phishing email, a data point from the experiment forms a tuple $(i, p_i)$ where $i$ is the round number and $p_i$ is the click-through rate. Using a simple linear regression technique, we compute the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ for line $\hat{p_i} = \hat{\beta}_0 + \hat{\beta}_1 * i$ that best fits all data points of the phishing email. Based on this regression line, we estimate $p_1$, the value of click-through rates in the first round (i.e., before any training is given). This click-through rate is what we refer to as *email persuasiveness*. This takes into account the bias of the round in which the emails are given.

For example, the "Order" email was given in the first, second, and sixth rounds with click-through rates of 28.5%, 36.3%, and 10.7%, respectively. Using regression, we estimate the email's persuasiveness for average users in the first round, before any training (32.5%). Table 3 shows the estimated email persuasiveness for all the emails in the dataset.

| Email | Score | Email | Score |
|---|---|---|---|
| Celebrity | 0.2 | AGCE | 6.4 |
| Sports | 0.3 | Federal | 6.4 |
| Newsletter | 0.5 | Domain | 7.1 |
| Dragon | 1.0 | Charity | 7.6 |
| Big Box | 1.5 | IQ Test | 7.8 |
| Bank 2 | 2.4 | Tax | 8.0 |
| Bank 1 | 2.7 | Funds | 8.5 |
| Warehouse | 3.0 | Password | 9.3 |
| Certify | 3.5 | Cellular | 9.8 |
| NACA | 4.0 | Complaints | 25.3 |
| Malware | 4.5 | Fax 2 | 27.3 |
| Outfitters | 4.5 | Fax 1 | 28.1 |
| Secure Mail | 5.6 | Order | 32.5 |
| Bazaar | 6.2 | Shipping | 43.9 |

**Table 3:** Email persuasiveness computed by linearly regressing over the adjusted click-through rates. The email persuasiveness shows the click-through rate of untrained average user.

## 5 Normalization Methods

To enable more accurate assessment of participant performance, we normalize and rescale the click-through rates for each data point (i.e., a phishing email in one round to a certain group of employees). To find an appropriate normalization methods, we have evaluated a list of possible general normalization methods:

**Z-score.** The standard score, or Z-score, is the signed number of standard deviations by which an observation or data is above the mean [1]. In this method, we use the persuasiveness score of the emails as the mean value $\mu$. To normalize a click-through rate $x$ for an email with mean $\mu$ and standard deviation $\sigma$, we compute $\frac{x-\mu}{\sigma}$.

**M-ratio.** This method scales the click-through rates on a link inside a phishing email linearly, relative to the highest (i.e. Maximum) click-through rate on that type of email. Scaling is done by dividing the click-through rate to the maximum value multiplied by 100. This gives

a percentage between 0 and 100.

**P-ratio.** This method scales the click-through rates on a link inside a phishing email linearly relative to the email persuasiveness computed in the previous section. Rescaling is done by dividing the click-through rate by the corresponding persuasiveness score of email from Table 3. This is the only method among others listed here, that uses estimated persuasiveness of phishing emails computed based on linear regression.

**L2-Norm.** In this method, click-through rates of each email type are rescaled to form a vector with norm 1. For click-through rate x of a given email type, we use the following formula for normalizing the click-through rates: $\frac{x}{\sqrt{\sum_{i \in emailtype} x_i^2}}$.

None of the listed methods are being proposed or used in the previous studies of phishing trainings.

**Goodness Metric.** To determine which method is more suitable for normalization, we defined a "goodness" metric. The idea is that an effective normalization method would show a monotonic trend for the click-through rate for groups, unlike, for example, what we observed in the raw click-through rates. The assumption of the monotonic trend is valid regardless of the effectiveness of the phishing exercise, as the click-through rate remains constant if the training is not effective, and decreases otherwise, if emails have similar persuasiveness. Therefore, a rescaling method is preferred that better represents the monotonicity of data points of each group. We checked the monotonicity of the data, by fitting a line with the linearly least square distance from the rescaled data-points. We use R-squared ($R^2$), a statistical measure of the closeness of the points to the fitted line to asses the monotonicity of the rescaled click-through rates for each group. The goodness metric $\lambda$, then is the average monotonicity score $R^2$ computed for all groups:

$$\lambda = \sum_{i=1}^{N} \frac{R_i^2}{N}$$

Where $N$, number of groups, is 32 in this dataset. $R_i^2$ is the R-squared for linear-regression over normalized scores of group i.

Table 4 shows the goodness score $\lambda$ for each normalization technique. The $\lambda$ for raw click-through rate is also given for comparison. As it is observed, the P-ratio performs best in this instance. Therefore, we use the P-ratio method for normalizing the click-through rates.

# 6 Evaluating Exercise Data

In this section, we use our methods from the previous section to perform an improved evaluation of click-through rate trends and the effectiveness of training.

| Normalization Scheme | Goodness ($\lambda$) |
|---|---|
| Raw Click-through | 0.27 |
| P-ratio | 0.43* |
| M-ratio | 0.34 |
| Z-score | 0.39 |
| L2-Norm | 0.40 |

**Table 4:** Goodness score $\lambda$ for different methods of normalization, computed based on the average of $R^2$ of regressed lines. P-ratio performs better than other methods.

## 6.1 Overall Click-Through Rates

Most organizations are primarily interested in trends in the overall click-through rates, as it presumably indicates the effectiveness of their embedded phishing training in reducing the susceptibility of their employees to these attacks. The easiest way of measuring this rate in each round is to compute the percentage of participants who click on phishing links. Applying this approach to our study data, one may conclude that the overall click-through rate only decreased from 7.9% in the first round to 5.13% in the last round of the phishing exercise (See Table 1), a drop that is not particularly significant. This computational method may also give the impression that employees of the company are protected, since a 7.9% click-through rate is quite impressive (compared with the click-through rates reported in [4, 14, 20]). However, a closer look at the data shows that the click-through rate has not changed uniformly for all types of emails. For example, click-through rate on the "Order" email in the first round was 28.5%, and decreased to 10.7% in the last round. In comparison, the click-through rate of the "Certify" email changed from 3% in the first round to 0.8% in the last round. We conclude that solely considering the raw click-through rates would not allow the company to fairly interpret the observed trend. As we analyzed in section 5, the P-ratio suits better for normalizing the click-through rates.

Using the P-ratio normalization method, we normalized the click-through rates and then, computed the overall click-through rate for each round. This was done by averaging the normalized click-through rates for all phishing tests in that round. Figure 2 shows the overall click-through rate in different rounds computed in this way. For comparison, the figure also shows the average raw click-through rates in different rounds. According to the P-ratio, the susceptibility (i.e., click-through rate) of the employees has decreased from 80% in the first round to 40% in the last round. Moreover, the normalized click-through trends show a steady and significant improvement over the course of the exercise.

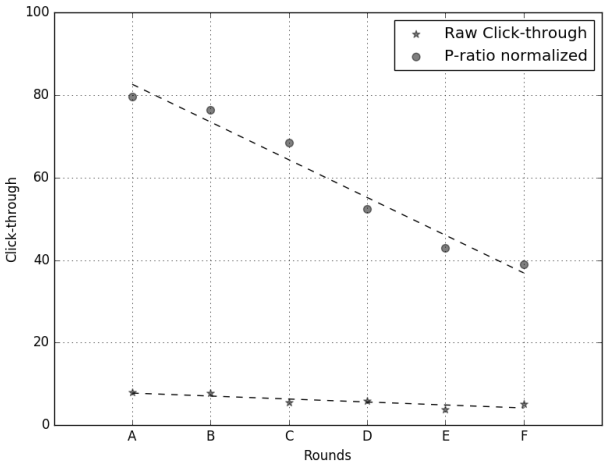


**Figure 2:** This figure shows the raw and P-ratio normalized click-through rate over the course of six rounds. According to the normalized value, the click-through rates have decreased from 80% to 40% as a result of phishing exercise.

## 6.2 Measuring Training Effectiveness

Organizations are interested in determining whether their training approaches are effective. The training method of the dataset under analysis used a few slides on how to detect phishing emails, followed by an interactive simulated email client (Inbox) with various emails (both normal and phishing). Subjects had to identify a certain number of suspicious elements to complete the training. This method of the training was also longer (4 minutes on average) than the training strategies of previous studies (which at maximum was one minute). In this section, we provide metrics to evaluate the effectiveness of training and detail our findings.

**Recidivism.** The mere fact that the overall performance of the participants has improved as a result of phishing exercises does not validate the effectiveness of the training. This is especially true when a study, such as the one for this dataset, is done in an actual workplace setting where interfering external factors are not controlled. Therefore, instead of considering the overall click-through rate, we decided to compare the percentage of the decrease in recidivism for trained vs. untrained participants. If the training is effective, it should be much less likely that trained participants will fall for a phishing message.

We study the effectiveness of the training by comparing the percentage of recidivism for users who are not-notified ($\frac{|Recidivist|}{|Not-notified|} \times 100$) vs. people who are trained ($\frac{|Recidivist|}{|Trained|} \times 100$). The result shows that 25.94% (N=2606) of not-notified participants fall for phishing schemes. In comparison, only 15.57% (N=2350) of the trained users fall for phishing, that is considerably lower than the not-notified participants. Fisher's Exact Test shows that the difference between these two groups is statistically significant ($p - value < 0.01$). This means that the embedded phishing training improves the resilience of the users to phishing.

| Persuasiveness | Previously Trained | Previously Not-notified |
|---|---|---|
| $P_1$ | 2.04% | 2.25% |
| $P_2$ | 4.67% | 7.69% |
| $P_3$ | 16.88% | 27.22% |

**Table 5:** The effect of previous training on emails with different level of persuasiveness.

**Persuasiveness and Training.** Previous work did not look at whether training is equally useful for phishing emails of different levels of persuasiveness. To understand this, we computed the click-through for three classes of phishing emails: unpersuasive ($P_1$), which includes emails with a persuasiveness level below 5%, moderately persuasive ($P_2$), which includes emails with persuasiveness levels between 5% and 20%, and persuasive ($P_3$) for emails with higher persuasiveness levels. For each class ($P$), we computed the probability of participants falling for such email after being trained in the previous round ($\frac{|Recidivist|}{|Trained|} \times 100$). For comparison, we also computed the probability for participants who were not notified after the previous round. Table 5 shows the computed click-throughs. The results show that training makes a more significant difference for email types that are initially more persuasive (i.e., $P_3$). In comparison, the improvement on click-through rate of less-persuasive phishing emails is not significant (i.e., $P_1$). This is possibly because primarily highly susceptible users fall for unpersuasive phishing emails, and it might be more difficult to educate this type of subjects.

## 7 Discussion and Future Work

**Identifying Other Sources of Bias** We focused on normalizing the effects of test phishing email persuasiveness and round for our improved evaluation metric. While these effects are probably some of the major ones that bias evaluation of these exercises, it is likely that there are other factors that might influence the accuracy of an evaluation metric. For this reason, we do not claim that our evaluation metric is optimal, but rather that it is likely an improvement over prior work on evaluating the effectiveness of embedded phishing training exercises.

**Improving Embedded Phishing Exercises** Our analysis shows a few avenues of improvement for embedded phishing exercises. We found that educating employees who fall for unpersuasive phishing emails, using the method of this campaign and potentially similar methods, did not demonstrate improved resilience to unper-

suasive phishing in subsequent rounds. The other result we found is that training participants over persuasive phishing emails significantly improves the average resilience to persuasive phishing emails. Therefore, using this sort of phishing training might better suit to educate average users, but not very susceptible ones. The experimental data that we have was not directly designed to explore this phenomenon. As future work we will design a set of experiments that will enable a better understanding of how different types of training might reduce a users susceptibility to unpersuasive phishing emails.

**Transferring Designs and Methods to Industry** As future work, we plan to transfer our methods for evaluating the efficacy of these exercises to the company that provided this dataset, as well as to other companies with which we have an existing relationship.

## 8 Conclusion

We have undertaken the analysis of embedding phishing results from a medium sized company. Our study presents methods to isolate and normalize key biases: persuasiveness of a test phishing email and the effects of which round a test phishing email was received.

Using our methods, we find that the improvement from training seems to be limited to more persuasive phishing emails and that there is no improvement for unpersuasive phishing emails. Based on our findings we can recommend improvements in the design of embedded phishing exercises that will likely increase their efficiency and effectiveness. We will release all of our data and tools so that others can improve upon our methods. The objective being to establish better embedded phishing training exercises design standards that will in turn cause companies to be more resilient to phishing attacks.

## Acknowledgments

The authors thank Vern Paxson and the reviewers for their helpful feedback. This work was supported in part by the National Science Foundation under contract 1619620 and a gift from Google. The opinions in this paper are those of the authors and do not necessarily reflect the opinions of any funding sponsor.

## References

[1] Transformed scores - standard scores. http://mypages.valdosta.edu/mwhatley/3900/standardized.pdf.

[2] ARACHCHILAGE, N. A. G., AND COLE, M. Design a mobile game for home computer users to prevent from "phishing attacks". In *i-Society* (2011), p. 485489.

[3] BAILEY, J. E., AND PEARSON, S. W. Development of a tool for measuring and analyzing computer user satisfaction. *Management Science 29*, 5 (1983), 530–545.

[4] CAPUTO, D. D., PFLEEGER, S. L., FREEMAN, J. D., AND JOHNSON, M. E. Going spear phishing: Exploring embedded training and awareness. *Security & Privacy, IEEE 12*, 1 (2014), 28–38.

[5] CHA, M., HADDADI, H., BENEVENUTO, F., AND GUMMADI, P. K. Measuring user influence in twitter: The million follower fallacy. *Icwsm 10*, 10-17 (2010), 30.

[6] COLETTA, V. P., PHILLIPS, J. A., AND STEINERT, J. J. Interpreting force concept inventory scores: Normalized gain and SAT scores. *Physical review special topics-physics education research 3*, 1 (2007).

[7] DODGE, R. C., CARVER, C., AND FERGUSON, A. J. Phishing for user security awareness. *Computers & Security 26*, 1 (2007), 73–80.

[8] HAKE, R. R. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics 66*, 1 (1998), 64–74.

[9] JANSSON, K., AND VON SOLMS, R. Phishing for phishing awareness. *Behaviour & Information Technology 32*, 6 (2013), 584–593.

[10] KREBSONSECURITY. Target Hackers Broke in Via HVAC Company. http://krebsonsecurity.com/2014/02/target-hackers-broke-in-via-hvac-company/.

[11] KUMARAGURU, P., CRANSHAW, J., ACQUISTI, A., CRANOR, L., HONG, J., BLAIR, M. A., AND PHAM, T. School of phish: a real-world evaluation of anti-phishing training. In *SOUPS* (2009), p. 3.

[12] KUMARAGURU, P., RHEE, Y., ACQUISTI, A., CRANOR, L. F., HONG, J., AND NUNGE, E. Protecting people from phishing: the design and evaluation of an embedded training email system. In *SIGCHI* (2007), ACM, pp. 905–914.

[13] KUMARAGURU, P., RHEE, Y., SHENG, S., HASAN, S., ACQUISTI, A., CRANOR, L. F., AND HONG, J. Getting users to pay attention to anti-phishing education: evaluation of retention and transfer. In *eCrime researchers summit* (2007), ACM, pp. 70–81.

[14] KUMARAGURU, P., SHENG, S., ACQUISTI, A., CRANOR, L. F., AND HONG, J. Lessons from a real world evaluation of anti-phishing training. In *eCrime Researchers Summit* (2008), IEEE, pp. 1–12.

[15] KUMARAGURU, P., SHENG, S., ACQUISTI, A., CRANOR, L. F., AND HONG, J. Teaching johnny not to fall for phish. *ACM TOIT 10*, 2 (2010), 7.

[16] LIAO, Q., AND LUO, X. The phishing hook: issues and reality. *Journal of Internet Banking and Commerce 9*, 3 (2004), 1.

[17] LIU, G., XIANG, G., PENDLETON, B. A., HONG, J. I., AND LIU, W. Smartening the crowds: computational techniques for improving human verification to fight phishing scams. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, p. 8.

[18] MAYHORN, C. B., AND NYESTE, P. G. Training users to counteract phishing. *Work 41*, Supplement 1 (2012), 3549–3552.

[19] MOHEBZADA, J. G., EL ZARKA, A., BHOJANI, A. H., AND DARWISH, A. Phishing in a university community: Two large scale phishing experiments. In *IIT* (2012), IEEE, pp. 249–254.

[20] SHENG, S., HOLBROOK, M., KUMARAGURU, P., CRANOR, L. F., AND DOWNS, J. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *SIGCHI, 2010* (2010), p. 373382.

[21] SHENG, S., MAGNIEN, B., KUMARAGURU, P., ACQUISTI, A., CRANOR, L. F., HONG, J., AND NUNGE, E. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *SOUPS* (2007), p. 8899.

[22] VERISON. 2017 Data Breach Investigations Report. http://www.verizonenterprise.com/verizon-insights-lab/dbir/2017/.

[23] ZIELINSKA, O. A., TEMBE, R., HONG, K. W., GE, X., MURPHY-HILL, E., AND MAYHORN, C. B. One phish, two phish, how to avoid the internet phish analysis of training strategies to detect phishing emails. In *HFES* (2014), vol. 58, SAGE Publications, pp. 1466–1470.