Understanding the Efficacy of Phishing Training in Practice

Grant Ho[◇†] Ariana Mirian[◁†] Elisa Luo[†] Khang Tong[⋆‡] Euyhyun Lee[⋆‡]
Lin Liu[⋆‡] Christopher A. Longhurst[⋆] Christian Dameff[⋆] Stefan Savage[†] Geoffrey M. Voelker[†]

[†]UC San Diego [◇]University of Chicago [⋆]UC San Diego Health

***Abstract*—This paper empirically evaluates the efficacy of two ubiquitous forms of enterprise security training: annual cybersecurity awareness training and embedded anti-phishing training exercises. Specifically, our work analyzes the results of an 8-month randomized controlled experiment involving ten simulated phishing campaigns sent to over 19,500 employees at a large healthcare organization. Our results suggest that these efforts offer limited value. First, we find no significant relationship between whether users have recently completed cybersecurity awareness training and their likelihood of failing a phishing simulation. Second, when evaluating recipients of embedded phishing training, we find that the absolute difference in failure rates between trained and untrained users is extremely low across a variety of training content. Third, we observe that most users spend minimal time interacting with embedded phishing training material in-the-wild; and that for specific types of training content, users who receive and complete more instances of the training can have an increased likelihood of failing subsequent phishing simulations. Taken together, our results suggest that anti-phishing training programs, in their current and commonly deployed forms, are unlikely to offer significant practical value in reducing phishing risks.**

1. Introduction

This paper focuses on simple, yet practically important, questions: what is the real-world efficacy of phishing training as practiced in the healthcare sector today and can we characterize the underlying reasons for these results?

The motivation for these questions is clear. By any measure, phishing remains one of the principal unsolved attack vectors in modern organizations. In spite of 20 years of research and development into malicious email filtering techniques, a 2023 IBM study identifies phishing as the single largest source of successful breaches (16% overall) [20]. This threat is particularly challenging in the healthcare sector where targeted data breaches have reached record highs. In 2023 alone, the US Department of Health and Human Services (HHS) reported over 725 large data breach events, covering over 133M health records, and 460 associated ransomware incidents (more than one per day) [2], [11].

Absent an effective technical defense, organizations have turned to security training as a means to staunch the bleeding. Our own institution admonishes each of us to "Be a Human Firewall" — to identify and resist enticements to click on suspicious email-borne links. Indeed, in many sectors it has become standard to mandate both formal security training on an annual basis *and* to engage in unscheduled phishing exercises in which employees are sent simulated phishing emails and then provided "embedded" training if they mistakenly click on the email's links [29]. Healthcare is no exception, and HHS recommends that all medium and large US healthcare organizations engage in both annual awareness training as well as monthly "simulated phishing and social engineering campaigns" [10].

The value of such training seems intuitive in the abstract, and has been justified by initial lab studies and modest-scale experiments demonstrating positive results. However, recent large-scale empirical measurements have brought these findings into question. Notably, the largest study of its kind — Lain et al.'s 15-month post-mortem analysis of embedded phishing training involving 14,000 corporate employees — found no positive effects from training (and even some evidence of a negative effect) [28].

In this paper we further explore this question, in the particular context of the healthcare setting, using data from a carefully designed quality-improvement effort at UC San Diego Health, a large healthcare institution we abbreviate as "UCSD Health". Critically, this dataset, covering 19,000 healthcare workers over 8 months, was meticulously designed to include explicit control groups (i.e., employees receiving no training), randomized assignment into different training conditions and phishing lures, and detailed analytics of training engagement and completion. Together, this design provides unusually rich evidence for investigating questions of training efficacy and allows us to make the following findings:

- *No clear benefit from annual security training.* We demonstrate no correlation between how recently a user in our study has completed annual "awareness" training and whether the user clicks on links in simulated phishing messages (§ 4.2).
- *Limited benefit from embedded phishing training.* Using randomized controlled trials and statistical modeling,

[◁]Currently a senior security researcher at Censys.
[‡]Collaboration through the Biostatistics, Epidemiology and Research Design (BERD) center's statistical consulting program in UCSD's Altman Clinical and Translational Research Institute (ACTRI).

embedded training provides a statistically-significant reduction in average failure rate, but of only 2% (§ 5.2).
- *The necessity of control groups.* Our data shows that while banal phishing lures may only attract clicks from 1–2% of users, other lures achieve upwards of 30% failure rates, far outstripping the potential benefits attributable to training (§ 3.2). Thus, training outcome assessments must always be couched in terms of an underlying control group that receives no training (something rarely included in research or practice).
- *Why phishing simulations fail to deliver training.* We show that phishing simulations fail to deliver appreciable training for two reasons in practice. First, only a small fraction of users "fail" in any given simulation (a median of 10%) and thus in any exercise the vast majority of users receive no training. Moreover, failing to click on a lure is not strongly predictive of future outcomes and the majority of users (over 56%) clicked on a phishing link at some point in our study, whether or not they had received training. Second, those users who receive training typically fail to engage with training materials. By measuring "time on page" for embedded training materials, we show that over half of all training sessions end within 10 seconds and less than 24% of users formally complete the training materials (§ 6.2).
- *The relative value of training type.* Our data includes separate cohorts that vary in the nature of the training content (generic anti-phishing information vs. training contextually related to the email received) and the mode of delivery (static webpages vs. training requiring interactive engagement). For the subset of users who both receive training *and* engage with the training materials to completion, we show significant differences in relative outcome. While static training results are negatively correlated with future outcomes, we show that interactive training can reduce the likelihood of clicking on a subsequent phishing lure by 19%. However, this latter result is small in absolute magnitude (likely due to the low completion mentioned earlier) and it remains an open question if this represents a selection effect or would generalize to the remaining users if they could be convinced to engage with training (§ 6.2).

Taken together, and in the context of other real-world controlled evaluations [7], [28], our results offer a sobering picture on the efficacy of existing phishing training. As currently designed and deployed in practice, training is unlikely to offer significant value relative to its considerable expense in time and effort.

2. Background

Organizations routinely deploy cybersecurity training with the hope of improving employee security behavior, and to satisfy regulations or insurance guidelines [9], [13]. Our work studies the efficacy of two widespread types of training: (annual) security awareness training and embedded phishing training.

In annual "cybersecurity awareness" training, employees receive a mandatory training program, typically via an online training website, that aims to teach users a broad array of basic security best practices and threats to keep in mind. UCSD Health uses material from KnowBe4 for its annual training, which consists of a website that walks through a set of instructional videos as well as interactive question-answer quizzes about security threats like phishing.

During embedded phishing training [23], an organization periodically sends simulated phishing email messages to its employees. If an employee fails a simulated attack (e.g., clicks on an embedded phishing link), they are immediately redirected to a training website that notifies the user they fell for a phishing simulation and provides them with educational material about identifying phishing attacks. This form of training remains a widely deployed practice, with a multi-billion dollar industry providing this training as a service (e.g., Proofpoint, Barracuda Networks, etc.). UCSD Health uses the Proofpoint platform for its embedded phishing training.

2.1. Related Work

In this section we focus on two relevant areas of prior work: cybersecurity awareness training and embedded phishing training.

Security Awareness Training: Prior work on security awareness training has largely involved lab settings with university students, and measured training efficacy by comparing users' pre vs. post-training performance on security quizzes and surveys. These studies largely conclude that security awareness training leads to a positive security impact. For example, studies have examined many different forms of awareness training, ranging from in-person and instructor led sessions to custom-built educational videos. These prior efforts show that after completing training, users have higher scores on security quizzes and/or have higher accuracy at identifying phishing vs. legitimate messages in line-ups of different email messages [4], [35], [42], [43], [46]. Similarly, at a more macroscopic level, Kweon et al. found a correlation between spending more time on security training and fewer cybersecurity incidents in a study of 7,089 organizations in Korea [26].

In contrast, one prior study by Back and Guerette, which involved 2,000 employees at a US research university, found that users who completed awareness training were more likely to click on a phishing email link than users who did not complete the training [3]. However, the study itself adds a caveat that its negative result might stem from several confounders, such as not truly randomizing users in the training vs. control group and that as much as six-months may have elapsed between when a user completed the training and when they received the simulated phishing email. Thus, although prior work demonstrates that various forms of awareness training improve user performance on security tests, it remains an open question whether commonly deployed versions of such training today (e.g., an-

nually assigned training delivered via websites) help protect employees against malicious email they receive in-the-wild.

In addition, this prior literature collectively suggests that the potential protection gained from training diminishes over time. For example, prior work shows that although users' results on security knowledge quizzes improves immediately after training, after four to five months, users' performance deteriorates and no longer exceeds their pre-training levels [4], [6], [35]. Other work by Zhang et al. [46] suggests that training has an even shorter protective "shelf life", where the knowledge and improvement from a single training session disappears after just one month.

Embedded Phishing Training: Most prior research suggests that embedded phishing training improves users' ability to identify and avoid phishing attacks [14], [21]. Among the earliest work, Kumaraguru et al. proposed the idea of embedded phishing training based on principles from learning science theory [25]. From these principles, the authors argue that training that involves learning-by-doing (in particular receiving a phishing email and making a real decision to fall for the attack or not) and immediate feedback (i.e., providing users with immediate training if they fail the simulation) should provide effective anti-phishing education. Multiple lab-based studies involving role-playing exercises indicate that embedded training does help users identify phishing attacks, that users do retain educational knowledge about phishing attacks for at least several days, that multiple rounds of training provide enhanced educational knowledge, and that personalization and the educational form factor (e.g., text vs. graphics, static vs. interactive content, and gamification) impact the efficacy of training [22], [24], [30], [36], [40], [45].

Additionally, several studies in real-world organizations also suggest that embedded phishing training can reduce users' susceptibility to phishing attacks [18], [30]. For example, Hillman et al. [18] and Kumaraguru et al. [24] compared users' performance on simulated phishing email messages over time, and found that users' failure rates decreased over time. However, these studies did not employ or compare the performance of users who received training against a true control group to capture whether the decreased failure over time results from potentially "weaker" phishing lures over time versus beneficial knowledge learned from the training material. Separately, Siadati et al. analyze eight months of embedded phishing training results from an enterprise with 19,000 employees who received randomly selected subsets of 26 phishing lures [41]. Because their study does not involve a randomized controlled design, they construct a regression model to estimate the efficacy of anti-phishing training and conclude that training does appear to improve users' ability to identify persuasive phishing attacks.

In contrast to this large body of prior work, a recent study by Lain et al. suggests that embedded phishing training actually leads to worse performance for users, when compared to users who receive no training [28]. Their study explores data from a randomized controlled experiment where employees at a large organization received 8 simulated phishing emails over 15 months. In their study, the training consisted of a static webpage with several paragraphs of educational text and a button to optionally start an additional training exercise. Among a number of results, this study's findings indicate that users who received embedded training failed phishing simulations at a statistically higher rate than users in a control group who received no training at all. In a subsequent study involving a smaller set of users, Lain et al. [27] find that embedded phishing training can lead to small improvements (decreases) in users' future failure rates. But their analysis also shows that users who receive training perform no better than users who receive only "deterrent emails" that warn of consequences for failing future phishing simulations, which suggests the major benefit of training comes from simply reminding users of the phishing threat, rather than educational content. Separately, an earlier study by Caputo et al. [7], which also involved a randomized controlled experiment with static training, found no significant difference in the failure rate of users in a control group versus those in a training group, based on three rounds of phishing simulations. Related, Gordon et al. [16] study whether phishing training can help decrease the phishing clickthrough rate of high-failure users (who failed five or more phishing simulations during a two-year timespan). Their study used quasi-embedded training, where high-failure users received a separate email requiring them to take an intensive offline anti-phishing training. Across five subsequent phishing simulations, these high-failure users continued to fail at a higher rate than other users in the organization after completing the offline training.

Apart from efficacy, prior work has also explored the potential costs incurred by embedded phishing training. On the positive side, responses from employees, who recently received phishing simulations and participated in voluntary interviews, suggest that users view these training exercises favorably [27], [38], [39]. Users report that the phishing simulations provide them with opportunities to learn and test their knowledge, as well as making sure they remain vigilant of phishing attacks. However, other studies highlight various burdens and risks ranging from increased stress to users, the monetary and time costs from procuring and deploying phishing simulations, a decrease in users' perceived efficacy, and security fatigue from too many warnings or reminders [5], [14], [36], [39], [44].

Given the ubiquity of embedded phishing training and the potential costs it imposes, an important question is whether this training is effective in practice and why such discrepancies about its efficacy exist in the literature. Are these contradicting results the product of subtle design decisions (e.g., the content and advice in the training material)? Or do these results emerge due to the use of randomized controls in a real-world setting versus earlier work's large reliance on lab-based studies or non-randomized trials in real-world organizations? One goal of our work aims to provide some empirically-backed explanations that reconcile these conflicting results.

3. Methodology

Our study analyzes the performance of nearly 20,000 full-time employees at UCSD Health across eight months of simulated phishing campaigns sent between January 2023 and October 2023. UCSD Health is a major medical center that is part of a large research university, whose employees span a variety of medical roles (e.g., doctors and nurses) as well as a diverse array of "traditional" enterprise jobs such as financial, HR, IT, and administrative staff. For their email infrastructure, UCSD Health exclusively uses Microsoft Office 365 with mail forwarding disabled. On roughly one day per month, UCSD Health sent out a simulated phishing campaign, where each campaign contained one to four distinct phishing email messages depending on the month. Each user received only one of the campaign's phishing messages per month, where the exact message depended on the group the user was randomly assigned to at the beginning of the study (§ 3.1). In total these campaigns involved ten unique phishing email messages spanning a variety of deceptive narratives ("lures") described in Section 3.2. All of the phishing lures focused on drive-by-download or credential phishing attacks, where a user failed the phishing simulation if they clicked on the embedded phishing link.

3.1. Experiment Design

Annual Security Training: At UCSD Health, each employee must complete a standalone security awareness training once per year (with the material designed by KnowBe4). Since employees engage in this annual training concurrently but independently of the embedded phishing simulations, the two activities provide an opportunity to analyze the correlation between how recently employees completed their annual training and their performance on the simulated phishing messages sent during this study.

When employees first join, the HR system automatically assigns an employee this annual security training to complete within a few weeks. Once a user has completed their training, the system automatically reassigns this training to the user after one year (365 days) has elapsed. This training appears in the institution's standard HR learning platform (e.g., similar to a standard enterprise HR system like Workday). Thus, throughout the duration of this study, the system automatically assigns annual security training for a rolling subset of the employees at UCSD Health.

Embedded Phishing Training: The procedure for embedded phishing training followed the standard process described in Section 2, where UCSD Health uses Proofpoint as its anti-phishing platform. For each simulated phishing message, if a user failed the phishing exercise (i.e., clicked on an embedded phishing link), the resulting webpage displayed anti-phishing training material and notified the user that they had fallen for a simulated phishing email. As with all embedded training, only users who failed the phishing simulation received the training content that month, since users who avoided the phishing URL would not click and load the training website.

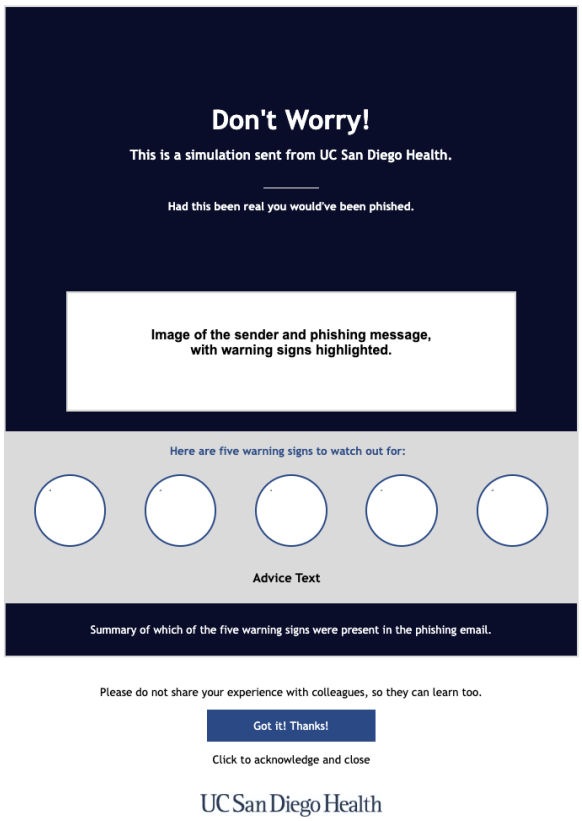


Figure 1: An example of the presentation and layout of the "contextual static" training content displayed to users (§ 3.1), shown as a template. Due to the terms of UCSD Health's contract with Proofpoint, we are unable to include images or examples of specific phishing messages or training material derived from Proofpoint's platform.

As part of their phishing exercises, UCSD Health explored five different training formats. As described below, this setup included a control group and four types of training that varied in their interactivity and whether the training content was contextualized/customized specifically for the phishing lure the user failed. These variations in training style help explore prior claims that interactive educational content and content related to a user's current situation and actions can help improve their learning experience [23], [37]. Figure 1 shows an example layout of one training ("contextual static") webpage.

1) Control Group: Users in this training group received no training material for the duration of the study. Instead, if a user failed a phishing simulation, the resulting webpage loaded a 404 ERROR message that did not mention anything about phishing or provide educational content. By comparing the performance of users in this control group against those in training groups, we can assess whether training improves the ability of users to avoid phishing attacks.

2) Generic Static Group: Users in this training group received a static educational webpage, taken directly from Proofpoint's embedded training library, that provides tips on how to avoid phishing attacks.
3) Generic Interactive Group: Users in this group also received a training webpage directly from Proofpoint's training library, that displays an example phishing email and walks users through an interactive question-and-answer training exercise with tips for spotting phishing attacks.
4) Contextual Static Group: This group received an adapted version of Proofpoint's static training webpage (used for the Generic Static group) modified so that the advice specifically related to the phishing lure the user received. Concretely, this training replaced the generic example phishing message with the actual phishing email the user had received, modified the advice content to mention which phishing warning signs and tips applied to the specific email, and used highlighting and a red-colored font to illustrate where these occurred in the displayed email.
5) Contextual Interactive Group: Finally, this group received a version of Proofpoint's generic interactive training webpage, modified to specifically relate to the simulation the user had failed. The example email was replaced with the exact phishing email the user received and the question-answer content was updated to accurately reflect whether a warning sign was present or not in the phishing email.

To analyze the efficacy of these different training programs, UCSD Health used a randomized, controlled study design whereby each user was assigned to one of these five groups for the entire duration of the study. Initially (and after cleaning the data, §3.2), each group consisted of roughly 3,950 users. But due to changes in employment status, the number fluctuates between 3,700 – 3,950 users per group during different months of the study. To help understand whether the sequencing or types of phishing lures a user received had any impact on user learning and future performance, each of the five training groups was further divided by randomly assigning users to one of four "tracks" (subsets), for a total of twenty distinct cohorts: five training groups each with four tracks. During each active month, a user received exactly one simulated phishing email, where the contents (lure) of the phishing email varied based on which track a user belonged to. For example, users in Cohort #1 (Control group, *Track 1*) received an account-related phishing email during Month 1; similarly, users in Cohort #2 (Generic Static group, *Track 1*) also received the same phishing email during Month 1. On the other hand, users in Cohort #6 (Control group, *Track 2*) received a document-related phishing email during Month 1. This design allowed us to expose distinct subsets of the Control and Training groups to different orderings of the phishing emails, which we also control for in our statistical analysis.

Power Analysis: To determine appropriate group sizes, we conducted a power analysis for testing for differences in phishing failure rates between control and treatment groups. We found that a size of N = 2,095 per group would be sufficient to achieve 90% power for detecting at least a 5% difference at a two-sided significance criterion of α = 0.05, for a two sample proportion test across a range of expected failure rates in the control group (between 10% and 50%). After adjusting for multiple comparisons by taking $\alpha^* = \alpha/3$ to compare three different types of training (interactive, static, and control), we would need N = 2,693 per group, which our study greatly exceeds.

Statistical Analysis: To determine whether our results show statistically significant relationships between training and users' performance on phishing simulations, we analyzed and fit multivariable generalized linear mixed effects (GLME) models to our study's data [15], [33]. GLME models allow us to compute an Odds Ratio (OR) that represents the relative change in a key outcome's value (e.g., a user's likelihood of failure) given a change in the main predictor variable (e.g., whether a user received training or not), while controlling for potential confounding variables (e.g., the different types of phishing lures and the number of times a user has previously failed a phishing simulation). In addition to an Odds Ratio, these models also provide a 95% confidence interval (CI) and a p-value (interpreted as statistically significant if $P < 0.05$) [15], [19].[1]

3.2. Data

During each simulated phishing campaign, UCSD Health collected the following anonymized information for each user via Proofpoint's simulation platform: whether the user viewed the email, whether the user failed the simulation (clicked the embedded phishing link), whether a failing user officially completed the embedded training (by clicking an "acknowledge" completion button at the end of training), the amount of time a failing user spent on the embedded training (in seconds), and the ISP and AS of the IP address that loaded the training webpage (if a user failed the simulation). Additionally, for each phishing simulation, UCSD Health augmented the data to include the number of days since a user last completed their annual security training and whether or not a user logged into Office 365 during the duration of the phishing campaign.

Data Cleaning: We took several post-hoc measures to ensure that our dataset contains only active, full-time employees; since inactive accounts will not open email messages, these data cleaning steps allow us to better estimate the true failure rate. First, UCSD Health security team members helped us exclude (at monthly granularity) users who were not active, full-time employees at the time based on HR data. Similarly, the data excluded users who had never been assigned annual training, which removed temporary users or email addresses corresponding to service/role accounts that

1. Although we cannot share any data from our study, we have released R code for running the final models and analysis in our paper: https://github.com/ucsdsysnet/phishing_training_code_oakland2025.

Phishing Lure	# of Users	Avg Failure Rate
Outlook Pwd	4,931	1.82%
Login Account	12,720	1.85%
Open Enroll	14,691	7.62%
Shared Doc (Microsoft)	15,683	8.99%
OneDrive Medical	18,438	9.20%
Docusign	23,526	9.63%
Building Evac	17,359	10.33%
Traffic Ticket	17,676	18.60%
Dress Code	4,954	27.65%
Vacation Policy	17,923	30.80%

TABLE 1: The number of recipients for each phishing email across our entire study and the average failure rate across all recipients. Appendix A contains more details about each email.

were accidentally included in the user list. As a final step, we also removed any user who had not logged into Office 365 during the entire duration of our study. In total after these cleaning steps, the data consisted of 19,789 active, full-time employees (with a random ID for each user).

Additionally, the security team at UCSD Health followed Office 365 and Proofpoint's best practices for ensuring that phishing simulations were not unintentionally disrupted or triggered by email protection settings. These steps included configuring specific rules ("allowlisting") in Office 365 for the phishing simulation domains and email content to prevent Office 365 from quarantining these emails and/or crawling the URLs. For one of the control cohorts in each of Month 3 and 4, the allowlisting rules were incorrectly configured and O365 quarantined nearly all of the phishing simulations for users in these cohorts. We conservatively excluded the data from all users in the affected cohort in each of these two months. Finally, during two of the early months in this study, one setting to prevent Office 365 from crawling URLs was not configured correctly for some phishing campaigns. As a result, we applied an additional data cleaning step where we removed any failure events where the ISP/AS of the clicking IP address corresponded to Microsoft entities. Thus, if both Microsoft's crawler and a real user clicked on the embedded phishing URL in an email message, our dataset would contain only the legitimate user's click (failure). However, if only Microsoft's crawler visited the embedded URL, our dataset would not report any failure for that user.

Summary Statistics: Table 1 shows the total number of users who received each of the ten phishing email lures across the duration of the study, and the total failure rate for each lure. The ten phishing messages were variations of five distinct lures (deceptive narratives), and specifically focused on email messages relevant to enterprise employees: two messages used a lure stating the recipient had some issue with their account or password, three lures attempted to deceive the user with a shared work document that the user needs to view, two lures discuss potential benefits policy changes (e.g., vacation or open enrollment), two other lures involved social protocol updates (e.g., dress code and building evacuation changes), and the final lure notified the user of a traffic / parking violation.

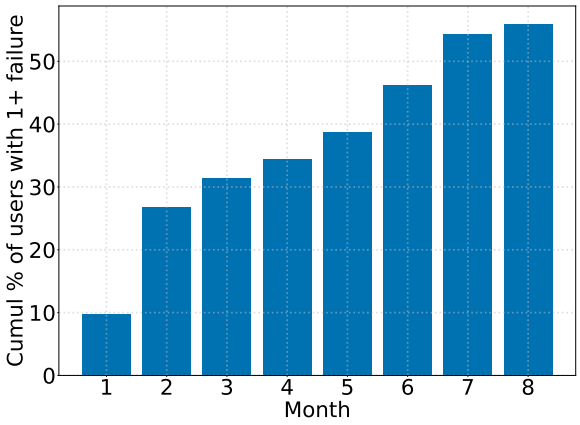


Figure 2: The cumulative percent of users who have failed.

Note that not all users received every phishing message (resulting in different numbers of recipients per phishing lure as seen in Table 1). The exact eight phishing messages a user received depended on their track (randomized order of phishing emails) as described earlier in Section 3.1. As a result, in some cases only one track received a specific phishing lure. However, every track (and thus every phishing email) contained a fixed subset of users in the control group and each of the four training groups, allowing our statistical analysis (§ 5.1) to accurately compare the efficacy of embedded training against users in the control group. Additionally, one of the tracks received the same phishing email ("Docusign") in Months 2 and 8 (increasing the number of recipients beyond the total number of users). We computed all our statistical models both including and excluding the Month 8 data for users with this "repeat" phishing simulation, and the results do not change.

Table 1 underscores that the efficacy (average failure rate) of the simulated phishing campaigns varies significantly among different phishing lures (from 1.8–30.8%), and even between messages that follow similar themes (e.g., 7.6% for open enrollment benefits vs. 30.8% for vacation policy). This variability illustrates the need for randomized, controlled comparisons and careful statistical analysis to evaluate whether a change in a user's phishing failure is due to training they received or other confounding factors, such as subsequently receiving "weaker" phishing messages that users can more easily identify.

User failure rates over time: Across all phishing simulations, 56% of users (11,077 out of 19,789) failed at least once by clicking on an embedded phishing URL. Additionally, 25.9% of users failed at least two phishing simulations, 9.8% failed at least three, and 3.5% failed at least four exercises; one user failed every single simulation.

Figure 2 shows the percent of users who failed at least one simulated phishing campaign over time. The results show a steady increase in the number of users who fail the phishing simulation throughout the study. In the first month 9.7% of the users clicked on a phishing URL, and after the eighth month 56% of users in the population

have clicked on at least one phishing URL. This steadily increasing failure rate suggests that with enough time and effort, attackers would likely be able to fool a large fraction of an organization's employees into falling for a phishing attack, and that users who avoid earlier phishing emails will not necessarily avoid future attacks.

3.3. Ethical Context

Similar to Lain et al. [28], the data for this study was collected as part of an existing and regular corporate practice of UCSD Health. This practice included both the requirement that all employees complete annual security awareness training (which includes specific material on defending against phishing) as well as active, unannounced, phishing simulations designed to test employees and deliver embedded training if appropriate. These activities were undertaken by UCSD Health IT staff, at the direction of their leadership, as part of an effort to ensure compliance with both internal best practice policy and external regulatory requirements. Both activities, i.e., mandatory annual training and periodic phishing simulations, were known to employees, but they were not informed precisely when the simulation exercises would take place, nor was there any post-exercise debriefing beyond the delivery of training content for those who triggered it. The study design in this work (i.e., the randomized controlled trials, different lures and training types) arose from a UCSD Health quality improvement initiative to improve anti-phishing training efforts and, ideally, reduce subsequent failures among employees. These efforts, and additional collaboration with researchers on this paper, received approval from UCSD Health's security, legal and research compliance teams. Users were also assured that their performance on the phishing simulations and training would have no bearing on their employment status.

As the non-UCSD Health authors of this paper became involved, they sought to collaborate and use this data to explore the broader research questions described herein. Prior to commencing any such work, we formally requested and received additional approval from UCSD's human subjects review board, which declared the study to be non-human subjects research as the data collection was pre-existing and the resulting data was anonymized. This subgroup worked with UCSD Health's IT team on anonymized data that only reflected whether each abstract user, randomly assigned to a particular cohort, clicked on a link in a given month and, if so, how long they stayed on the training page and whether or not they completed the training.

Given this context, and a consequentialist lens for evaluating harms, we believe our analysis posed minimal risks. Employees were already subject to the overhead of annual training and embedded phishing simulations as a byproduct of existing organizational policy at UCSD Health. Thus, our analysis of this data did not create any new workload or unique risk. Moreover, since individual employees were never identified, there was no particularized privacy injury (beyond that inflicted by the pre-existing training requirement itself). Balanced against this status quo, our study has the beneficial potential to understand the utility of such training as well as the factors that lead to its efficacy or inefficacy. Furthermore, in this case, the impact of either positive results (identifying effective training approaches) or negative results (finding that training offers no value and should be discontinued) have the clear potential to benefit participants.

4. Annual Security Awareness Training

As a baseline, we first explore whether user failures in the phishing simulations are correlated with the time since their last annual *security awareness* training. As part of an existing policy that predates our study, all full-time employees at UCSD Health must complete an annual cybersecurity awareness training program (§ 3.1), which uses material provided by KnowBe4, a third-party company that specializes in cybersecurity training. The content of this training focuses heavily on social engineering attacks, and includes short quizzes and anti-phishing advice; Figure 5 in the Appendix shows a sample screenshot from this training.

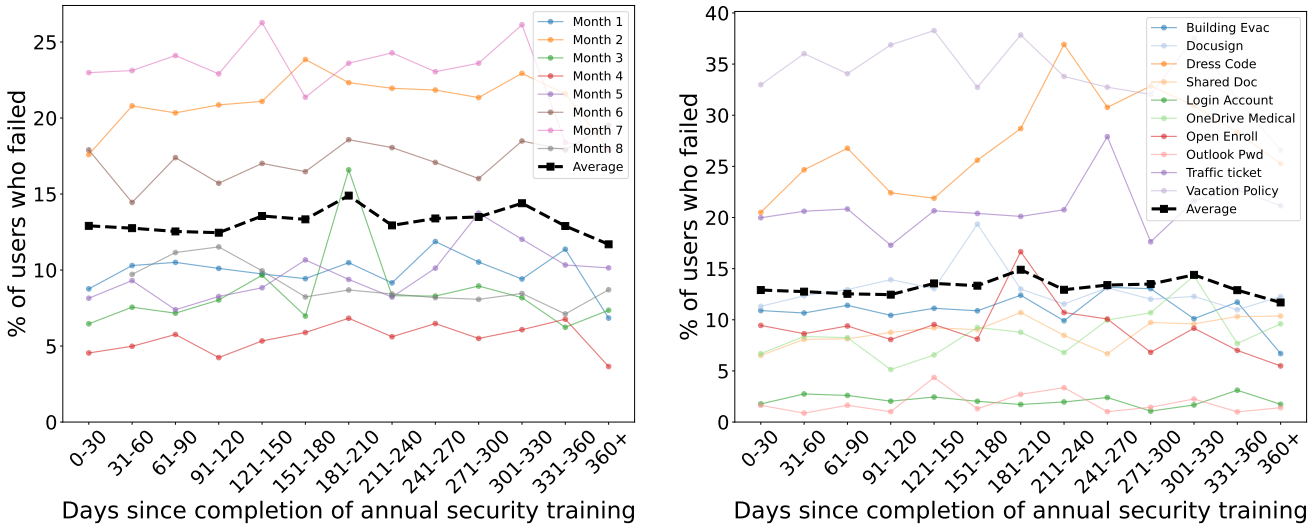
To understand whether this annual training correlates with better security outcomes, in this section we examine if employees who recently completed the training were less susceptible to the simulated phishing attacks we sent each month. Unlike embedded phishing training, where users can freely exit the loaded training webpage, this annual training is part of an organization's traditional HR system with institutional compliance policies. Thus it reflects a mandatory form of training that users must engage with. Ultimately, our analysis shows no association between how recently an employee completed their annual training and whether they failed a simulated phishing attack.

4.1. Data

For each month of a simulated phishing campaign, our data contains the completion status of each employee's annual training: the date the employee was assigned training and the date they last completed the security training. Across all eight months, an average of 83.7% of employees had satisfied the annual security training requirement at the time users received a phishing email. The remaining users are overdue on training (i.e., completed their most recent training over 365 days ago).

4.2. Results

Overall, we find no evidence that annual security awareness training correlates with reduced phishing failures. Specifically, based on the results of prior work (§ 2.1), we assume that if annual training provided anti-phishing knowledge, then employees who have recently completed training should have lower failure rates than users who took the training long ago. In particular, prior work has shown that users' scores on security knowledge quizzes improves shortly after taking awareness training, but that their performance reverts to their pre-training levels after a few



(a) **(b)**

Figure 3: Average phishing failure rate based on the number of days (in 30-day intervals) between when a user completed their annual security awareness training and when they received a phishing email each month. Figure 3a shows this temporal relationship based on user performance for each month in our study, and Figure 3b shows this relationship broken down for each phishing email lure. Although some fluctuations exist across different months and phishing lures, the overall trend (shown as the dashed, black average line) shows no strong relationship between how recently users completed their annual training and their failure rate with simulated phishing (§ 4.2).

months [35]. Because the first employment dates of users vary over the years, each month of phishing campaigns has a naturally-induced range of times for how recently users have completed their annual training. This distribution allows us to study whether users who have recently completed the awareness training have lower phishing failure rates.

We constructed a GLME model (§ 3.1) to examine the temporal relationship between failure rate and days since completion of annual security training. Our model treats the failure of a user for a given month as the key outcome, defined as a binary variable, and the main predictor variable is the number of days since completion of annual security training. We include the following confounders in our model:

- the phishing lure (categorical, e.g., "Login Account"),
- the order in which a user received phishing emails ("track": integer 1–4),
- seasonal differences (e.g., which month a user received a particular phishing message: integer 1–8), and
- how many times a user had previously failed (integer).

Additionally, we included a random intercept in our model to account for the repeated measures (multiple months of data) from each user. Ultimately, our model shows no significant association between the time since a user last completed training and their likelihood of failing a phishing simulation (OR = 0.998 per 30 day increase, 95% CI: (0.996, 1.000), P = 0.06).

Visually, Figure 3a shows the average phishing failure rate as a function of the number of days that have elapsed since users last completed their annual training (in 30-day intervals). The bold dashed line shows the average failure rates for all users combined. Reflecting the statistical results, the average failure rates for all users are independent of the number of days since employees completed their annual security training. As an example, consider the extreme ends of the time range: (i) users who have completed their annual training within a month before receiving a simulated phishing message, and (ii) users who are non-compliant and have not done their annual training for over a year. We observe no significant difference in phishing failure across all months. The lighter background lines show the proportion of employees who failed the simulated phishing for each month of the study. Different months have significantly different failure rates due to the nature of the lures used in the phishing campaigns for those months (Table 1). While individual months show some variations that average out when combined, the trends remain the same. Independent of the phishing lures used, failure rates are consistent across the times since completing the annual security training.

Figure 3b shows a similar graph, but the background lines show the failure fraction for each phishing lure in our study. Although a few individual phishing lures show some differences across the time ranges, the majority of phishing lures and the overall trend show commensurate failure rates over time. Furthermore, as discussed above, our GLME model finds no relationship between phishing failure and how recently users completed their annual training, while controlling for potential confounders that could account for the small fluctuations in a few of the lines.

5. Embedded Phishing Training

In this section we use statistical modeling to explore the extent to which embedded phishing training helps em-

Phishing Lure	Control	Generic Static	Generic Interactive	Contextual Static	Contextual Interactive
Login Account	3.44%	1.14%	0.97%	1.27%	1.13%
Outlook Pwd	1.62%	1.72%	1.85%	2.41%	1.52%
John Davis	9.56%	7.01%	6.4%	6.38%	7.45%
Docusign	11.06%	9.98%	10.05%	10.2%	9.75%
OneDrive Medical	9.89%	9.37%	9.25%	8.54%	9.16%
Open Enroll	9.02%	6.67%	7.01%	6.68%	6.76%
Vacation Policy	31.02%	30.58%	30.58%	31.99%	29.85%
Traffic Ticket	20.39%	20.07%	17.25%	16.37%	19.37%
Building Evac	11.67%	8.25%	8.55%	8.4%	9.32%
Dress Code	29.96%	27.01%	26.98%	27.41%	26.88%

TABLE 2: Failure rate for each phishing lure across all users in each training group (§ 3.1) who received the corresponding phishing lure. The average difference between the control group and the training groups across all phishing emails is 1.7%.

ployees avoid phishing attacks in practice at UCSD Health. We also use insights from our analysis to provide one explanation for reconciling seemingly conflicting results from prior studies. Based on the use of randomized controlled trials and our statistical models, we find that, in aggregate, users in the four training groups do have a statistically lower failure rate than users in the control group (§ 5.2). However, our analysis indicates that this security improvement is quite small: on average, users in the training groups have only a 1.7% lower failure rate than those in the control group, and for several phishing campaigns, at least 10% of users in every group failed the simulated attack.

5.1. Analysis Overview: Embedded Training

Intuitively, if any training group led to improved security knowledge and phishing avoidance, then users in that training group should outperform the control group's users across each of the phishing simulations. Table 2 shows the average failure rates for each training group on the ten different phishing campaigns in our study. As seen in this table, our empirical results paint a complex picture of embedded training's efficacy.

On one hand, each of the training groups outperform the control group across most of the phishing campaigns. For example, on the "Login Account compromise" phishing lure, most of the training groups have a 3–4$\times$ lower failure rate than the control group. However, for several phishing campaigns, the control group and training groups fail at equivalent rates; and, in some instances, the control group even has lower failure rates than a few specific training groups (e.g., for the "Outlook Password reset" phishing lure and the "Vacation Policy" phishing lure). Additionally, no single training group universally performs the best. For example, users who received the Generic Interactive training have the lowest failure rates on several phishing campaigns (e.g., "Login Account" and "John Davis"), but in other phishing campaigns, a different training group (sometimes even the control group) has the lowest failure rate (e.g., "Open Enrollment" and "Traffic Ticket").

To make sense of the overall picture, while controlling for potential confounders, we again fit a multivariable generalized linear mixed effect model [15] to assess whether any of the training programs provided improvements over the control group in reducing phishing failure. We exclude data from Month 1 from this model, since users do not see the embedded trainings until after failing a simulation (and all control and training groups have equal failure rates in Month 1). We control for the same confounders as in our previous model (§ 4.2), such as the specific phishing lures a user received each month. As shown earlier in Table 1, confounders like the inherent strength of different phishing lures might have a strong impact on a user's performance across months. Thus, in addition to dividing users across multiple tracks (different phishing lure orderings), statistically controlling for these confounders allows us to more accurately understand the relationship between training and phishing failure.

5.2. Results: Embedded Training

Based on the results of our multivariable GLME model, a user in a training group was 9.5% less likely to fail a phishing simulation than a user in the control group (OR = 0.905, 95% CI: (0.863, 0.950), $P < 0.001$). However, despite this statistically significant improvement, we emphasize that the overall magnitude of improvement is small. As shown in Table 2, in absolute terms across all users, the best performing training groups only have a 1–4% lower average failure rate than the control group for most simulated phishing attacks.[2] Moreover, we observe that all of these groups still have failure rates of over 15% for several phishing simulations, and that some phishing lures achieve over 25% click rates. Such failure rates significantly overshadow the improvements provided by the training: better phishing lures increase failure rates much more than the training approaches decrease them.

5.3. Differences from Prior Work

In the context of prior work, our results paint a more complicated picture about the efficacy of training. As discussed in Section 2, most prior studies have found that

2. Note that the odds ratio from our model reports the *relative* reduction in a user's likelihood of failure (9.5%), which translates to a 1.7% absolute difference in the full population's aggregated average failure rate.

Statistic	Generic		Contextual	
	Static	Interactive	Static	Interactive
Sessions w/ 0 sec	39.7%	51.3%	37.3%	44.3%
25th percentile	0 sec	0 sec	0 sec	0 sec
50th percentile	7 sec	0 sec	10 sec	6 sec
75th percentile	19 sec	24 sec	27 sec	48 sec
90th percentile	34 sec	70 sec	52 sec	101 sec

TABLE 3: Summary of the time spent on different embedded training sessions across all phishing campaigns.

phishing training can help users avoid future attacks. However, we note that these studies focused primarily on simulated lab environments with voluntary, non-enterprise users operating outside of their typical working environment. In contrast, studies such as [7], [27], [28] that employed controlled, randomized designs (which better account for confounding variables, such as differences in the inherent strength of phishing lures) suggest that training can make users more vulnerable to future attacks, or at best, provide small overall improvements to users' future failure rates.

Our study aimed in part to clarify this discrepancy, and ultimately aligns with results that question the efficacy of training. Although we observe a statistically significant improvement for users who receive training, the magnitude of this improvement is very small (on average a 1.7% lower failure rate). Like Lain et al. [28] and Caputo et al. [7], our study involves real-world users receiving the phishing simulations in-situ with a randomized controlled experiment design. Additionally, going beyond prior efforts, our statistical analysis explicitly accounts for a variety of potential confounding factors. This experiment design could explain why our results show only minimal benefits from training, as compared to much prior work. In the following section, we present analysis results that provide additional explanations for why such differences exist across the literature.

6. Embedded Training Engagement

Both earlier work and our work differ in several key conclusions about phishing training (e.g., some prior work shows large benefits to training [23], while others show no benefit or worse security outcomes [28]). This section aims to provide one explanation for these differences by understanding how variations in the design of different studies manifest in real-world user behavior and outcomes.

In particular, studies that show more negative results for training efficacy (e.g., our work and the work by Lain et al. [28]) often involve *enterprise users* receiving phishing emails and training in their *natural real-world* environment. Earlier work that pioneered the idea of embedded training evaluated its efficacy in lab settings with dedicated volunteers. In these simulated settings, users often spent several minutes reviewing training material (e.g., in one study users in the training groups spent 15 minutes participating in training [40]). However, in the real-world, employees are not specifically interested in receiving anti-phishing training and can simply exit out of the embedded training website.

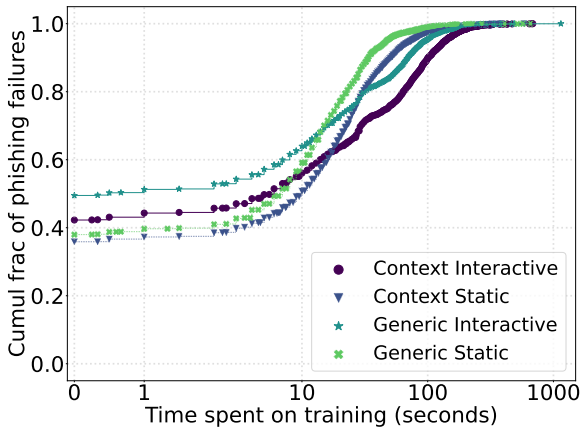


Figure 4: The distribution of the amount of time spent on each training program across all phishing training sessions (all instances of a training group user failing a phishing simulation).

To investigate this factor further, we analyzed the time users spend in-the-wild on embedded phishing training, a key variable whose impact is understudied in prior work. For this analysis, we exclude users in the Control group since they do not receive actual training content. Our results suggest that users in real-world settings spend very little time on the training websites, with significant portions of users leaving the training website immediately. Digging deeper, this lack of engagement does not appear to stem from training fatigue (from multiple months of phishing), but most likely stems from other factors (such as an inherent disinterest in prioritizing security as a computer task).

However, we find that for interactive training, users who fully complete a training session have a lower likelihood of failing future phishing exercises than those who do not. This same relationship does not exist for users who receive static training. In fact, consistent with recent work [28], users in static training groups who complete more training sessions have a higher likelihood of failing phishing simulations (§ 6.2). Collectively, these results suggest that organizations should carefully consider how users interact with embedded training in practice and the training content they deploy.

6.1. Do users engage with training in-the-wild?

Training time: Table 3 and Figure 4 summarize the distribution of how much time users spent on the embedded training websites across all training sessions (all instances of users failing a phishing simulation across all eight months). These statistics show that only a very small fraction of users engage with the training material. In the most extreme case, between 37–51% of all training sessions have no engagement at all: users simply close the page immediately.[3]

3. The training pages collect these measurements via Javascript, so it is possible that the presence of Javascript blockers could contribute to some of the training sessions that last 0 seconds.

Training	Completion Rate
Overall	24.0%
Generic Static	32.6%
Context Static	24.2%
Generic Interactive	15.6%
Context Interactive	23.5%

TABLE 4: The average completion rate of embedded training sessions across the study's duration (§ 6.1).

Furthermore, in over 75% of the training sessions, users spend less than 1 minute on the training page, and this proportion exceeds 90% for both static training groups.

These statistics suggest that, in practice, users spend very little (if any) time interacting with the educational material. Given this short duration, it is unclear how much educational value and defensive knowledge embedded security training provides, beyond increased paranoia or mere "contact exposure": the primary training users receive is simply the awareness that phishing exists and some implicit learning of the kind of phishing lure that they just failed. This small amount of time spent on the training material likely contributes to the small protective effect size in our results, and the similar average difference between the four types of training we studied. It also provides one explanation for why earlier work involving voluntary users in lab settings have much more positive results about the efficacy of training than more recent work studying training in real-world settings. For example, volunteer users in lab settings willingly and specifically spend time on training—potentially leading to better learning and outcomes—whereas the majority of users in-the-wild do not engage with the training material with similar interest or focus.

Training completion: Each training session includes a button at the end of the material that users can click to officially acknowledge and thereby "complete" the training (Figure 1 shows an example at the bottom of the webpage). As summarized in Table 4, users completed 24% of the training sessions across the duration of our study.

Similar to the fraction of users who immediately exit training, the users assigned interactive training content completed the training at much lower rates than those who received static training. Since the training completion button only appears at the end of the content, only users who spend the time to answer every question in the interactive training can successfully complete the training.

Training fatigue: Because this study analyzes the results of eight months of simulated phishing campaigns, the low user engagement with training might result from accumulating fatigue: as users see more training, they become less interested in engaging or completing the sessions. However, we found no evidence of this phenomenon.

Specifically, across our entire study, training group users failed 15,332 phishing exercises. Each failure produced a training session, where 9,049 sessions corresponded to the first time a user ever saw training and 6,283 sessions corresponded to "repeat" sessions, where a user had previously received at least one training session before. First-time training sessions had a completion rate of 23% whereas "repeat" training sessions had a slightly higher completion rate of 24%. Since fatigue should create a lower completion rate for repeat training sessions, the similar completion rates suggest that training fatigue is not the primary reason for low user engagement with the training material. Instead, the small amount of participation time and low completion rates likely occur because users view security as a secondary goal when using their devices (as prior work from usable security suggests more broadly [1], [12], [17]).

6.2. Training Engagement and Future Failure

Although many users spent very little time on embedded training, a small portion of users had greater engagement with the training material. For example, in 5.5% of training sessions, users spent over 90 seconds on the webpage. In this section, we present a statistical analysis that shows users who received interactive training have a lower likelihood of failing a future phishing simulation if they have previously completed a training session. However, this result does not hold for static training, and in the extreme cases, users who complete *multiple* prior static training sessions have a 15% higher chance of failing a future phishing exercise for each additional training they complete.

Similar to earlier analysis (§ 4.2 and § 5.2), we constructed six generalized linear mixed effects models to analyze the relationship between phishing failure and a user's engagement with training material. Specifically, we ran two sets of models, one set with and one set without an interaction term between training type (static vs. interactive) and a key user engagement variable. For each model set, we ran three models to examine the association between failing a phishing email and the following user engagement variables: (1) whether the user has officially completed a training in the past (binary: yes or no), (2) the total number of training sessions the user has previously completed (integer), and (3) the cumulative time users spent on training in the past (integer: in 30 second increments). In all six models, we control for the same set of confounding variables as we did in Sections 4.2 and 5.2. These models only compare data from users in the training groups, since users in the control group do not receive training content and thus do not have meaningful data about these key engagement variables, such as whether they "completed" the training.

Table 5 summarizes the statistical results from our two sets of GLME models, where each row corresponds to the key findings for one user engagement variable from our models. The "Overall" column corresponds to the results from our models that do not include an interaction term, and the last two columns ("Static" and "Interactive") show the results derived from the models that include an interaction between training type and the engagement variable. Odds ratios where the associated 95% confidence interval (CI) does not include 1.0 in its range are statistically significant with an α = 0.05 [15]. Collectively, these models show two interesting results: (1) only users who complete interactive

Model	Overall (n = 8831, 60486) OR	95% CI	Static (n = 4406, 30210) OR	95% CI	Interactive (n = 4425, 30276) OR	95% CI
1) Completed at least 1 Prev. Train.	0.963	(0.903, 1.027)	1.059	(0.978, 1.146)	0.809	(0.730, 0.896)*
2) Cumul. # Prev. Train. Completed	1.092	(1.034, 1.153)*	1.185	(1.110, 1.266)*	0.934	(0.856, 1.020)
3) Cumul. Time Train. (30s intervals)	1.008	(0.992, 1.025)	1.046	(1.017, 1.075)*	0.995	(0.977, 1.014)

TABLE 5: Summary of the key findings from each of our multivariate GLME models studying the association between a key training engagement variable and the likelihood of future phishing failures (§ 6.2). Only the results from our fully adjusted models are displayed (adjusted for number of previous failures, month, phishing lure, and track). Sample sizes are given as n = # of users, # of records. We ran Models 1-3 with main effects only (reported in the "Overall" column), as well as with an interaction between static vs. interactive training groups and the key engagement variables (the "Static" and "Interactive" columns report the results derived from this second model set). Results with an asterik indicate statistically significant odds ratios, with an $\alpha = 0.05$, since the 95% confidence intervals (CI) do not include a value of 1 within their range [15].

trainings have significantly better future performance, when compared to users who have received training but do not complete it (OR = 0.809, 95% CI: (0.730, 0.896)); and (2) users who have completed *multiple* static training sessions have an *increased* likelihood of failing a phishing exercise (OR = 1.185, 95% CI: (1.110, 1.266)).

Our data shows that an interactive training user had a 19% reduction in phishing failure rates if they previously completed the training (OR = 0.809, 95% CI: (0.730, 0.896)). On one hand, this statistical improvement from interactive training might indicate that completing the training provided useful security knowledge and education. On the other hand, this association might be the result of a self-selection bias, where some underlying differences exist between the population of users who choose to complete the training versus those who receive the training but do not complete it; in this latter case, the statistical improvement would correspond to a difference between these two user groups and not an effect of the training.

Our models provide some evidence that the improvement stems from the former reason (educational benefits). In particular, our models show no statistical improvement for users who have completed a static training session versus those who receive static training but refuse to complete it (OR = 1.059, 95% CI: (0.978, 1.146)). If the reduction in phishing failure rate was solely due to a self-selection bias, then we would expect to see a lower likelihood of failing for the users who choose to complete static training as well. However, our model does not show this result, which suggests that completing an interactive training may provide beneficial security knowledge to users, since self-selection bias alone does not correlate with improvement for other types of training.

Furthermore, our second model (row 2 in Table 5) shows that users who complete *multiple static* training sessions have a *18.5% increased* likelihood of failing for each additional training they complete (OR = 1.185, 95% CI: (1.110, 1.266)). In other words, completing multiple static training sessions correlates with increasingly *worse* security outcomes. However, we do not see a similarly negative outcome for users in the interactive training groups (OR = 0.934, 95% CI: (0.856, 1.02)). This increase in harm from this specific training style aligns with the findings by Lain et al. [28], which involved a training webpage similar to our static training sessions, where users in their study who received this type of training performed worse than users in a control group.

As before, this negative correlation may result from self-selection bias: for embedded training, only users who naturally fail multiple times have the ability to complete multiple training sessions. And, failing more times means these users have some characteristics that make them more susceptible to fall for phishing attacks than users who fail fewer simulations. However, the fact that we only observe an increase in harm for users who complete multiple static training sessions, but *not* for interactive training sessions indicates instead that differences in the training content itself may contribute to worse performance.

Collectively, our analysis suggests that completing interactive forms of training may provide additional improvements in the security knowledge of users (anti-phishing awareness), but static forms of training (e.g., an information webpage) do not provide the same benefit. Additionally, our models (Table 5) show no *beneficial* association between failure rate and completing multiple training sessions (OR = 1.092, 95% CI: (1.034, 1.153)) or spending more cumulative time on training (OR = 1.008, 95% CI: (0.992, 1.025)), which suggests that naively forcing greater training compliance and more time-on-page does not lead to a reduction in phishing susceptibility in many cases.

6.3. Training Engagement Summary

Altogether, the models and analyses from this section help explain why prior work observes such large discrepancies in the efficacy of embedded phishing training. In particular, our measurements show that very few users engage with embedded training in-the-wild, e.g., users complete only 15–24% of interactive training sessions (Table 4). This finding helps explain why some results, which involve volunteer users or lab settings with naturally higher training engagement, may have much more positive results about training efficacy.

Additionally, the content and style of training may help explain in part why some studies, such as Lain et al. [28], find a correlation between users who receive training and an increased (harmful) likelihood of failing future phishing exercises. In our study, for example, the small subset of users

who complete *multiple* static training sessions have a higher likelihood of failing future simulations than those who do not complete the training. The same negative correlation does not exist for users who receive interactive training. In contrast, the subset of users who fully complete one of these interactive training sessions have better future security outcomes than those who do not; however, only a small fraction of users actually complete an interactive training session and experience this increased benefit.

7. Discussion

Today nearly every enterprise mandates employee cybersecurity awareness training, and in particular anti-phishing training, as a defensive "best practice". The prevalence of these training efforts stems from both a long line of prior research arguing for its efficacy, combined with industry and regulatory pressures to codify verifiable compliance actions into best practice frameworks [31]. Unfortunately, when evaluated using data from real-world, large-scale randomized controlled trials, the promised benefit of these programs falls short.

First, we have shown that annual cybersecurity awareness training does not inoculate our organization against phishing attacks, nor does it provide significant protection towards that goal. Employees who recently completed such training, which has a significant focus on social engineering and phishing defenses, have similar phishing failure rates compared to other employees who completed their awareness training many months ago (§ 4). To the extent this result generalizes beyond UCSD Health, the lack of any meaningful impact on testable outcomes suggests that organizations should reconsider the value of this activity.

Second, our work highlights several operational challenges and subtleties in embedded anti-phishing training. Crucially, by its very design, only users who fail a phishing simulation receive embedded training. As a result, only a limited subset of an organization actually receives such training during each embedded phishing exercise. By providing no education for non-failing users, this design implicitly assumes that users who do not fall for one phishing lure do not need training to protect against future attacks. Unfortunately, our results show that the majority of users at our organization will eventually fall for a simulated phishing attack given enough time (§ 3.2). For example, hundreds of users who have successfully avoided seven prior phishing simulations eventually fail on the eighth simulated attack. These results suggest that embedded phishing training offers an inefficient means to educate users, nor does it accurately identify when and which users need training. This also suggests that studies with only one or a few phishing exposures very likely underestimate population-level susceptibility.

Finally, although our study finds that embedded phishing training correlates with a statistical reduction in subsequent clicks on phishing links, the size of this improvement pales in comparison to the efficacy of phishing attacks (§ 5.2). In particular, while some phishing lure only fool small numbers of users, nearly half of the phishing campaigns in our study convinced over 10% of their recipients to click on a link: more than double the 1–4% absolute reduction provided by training on different phishing lures. Thus, our results suggest that organizations like ours should *not* expect training, as commonly deployed today, to substantially protect against phishing attacks — the magnitude of protection afforded is simply too small and employees remain susceptible even after repeated training.

Reconciling Contradictory Results in the Literature: In the context of prior academic results, our work helps explain several seemingly contradictory findings. First, our work highlights the importance of conducting controlled, randomized comparisons to understand the efficacy of training. Because phishing exercises can have dramatically different success rates based on their message content (e.g., 2% vs. 30%), studies cannot simply compare a population's performance over time in isolation, since increases/decreases in failure rates may be entirely dominated by the strength of the phishing lures. This subtlety may explain why some prior work reports much higher efficacy of training than studies like ours that involve randomized, controlled trials.

Second, we provide insights on how much users actually engage with embedded phishing training in-the-wild, such as the time users spend on the training. In practice, the majority of the users in our study spend less than 30 seconds looking at embedded training content and less than one-third of users complete the training. This limited real-world engagement could help explain the difference between the negative results from our study and Lain et al. [28], when compared to prior work showing positive benefits from users in settings where training is supervised (e.g., in a laboratory experiment).

Third, our results also suggest that the potential benefits users can gain from training depend on its content and delivery method, but may incur diminishing returns. For example, similar to Lain et al. [28], we find that users who complete multiple *static* training sessions actually have worse phishing failure rates than those who do not complete the training. In contrast, users who complete interactive training sessions have a statistically lower likelihood of failing future phishing simulations (§ 6.2). However, regardless of modality, our data does not show any further reductions in failure rate from completing multiple training sessions.

Limitations: Our study's data comes from a single healthcare organization. Although this in-situ data spans over 19,500 users across a diverse set of job roles, the results from our data alone may not generalize to every organization and/or economic sector. However, our findings, questioning the efficacy of embedded phishing training, align with the results of other work that also used randomized controls [7], [28], which studied different organizations in different economic sectors with different embedded phishing training platforms and material.

For both of our analyses, on annual security awareness training and embedded phishing training, our control groups are not exact equivalents to users receiving no training. In the case of annual training, we did not have an explicit

control group (since users must take the training due to institutional policy). Nonetheless, the lack of any significant relationship between how long ago a user completed this training and their performance on simulated phishing emails suggests that the community should re-examine whether such training, as delivered today, provides meaningful security benefits. For our embedded phishing training, the control groups received an error webpage, which may allow some users to implicitly gain knowledge about phishing attack avoidance (by virtue of receiving a phishing email and some opaque feedback if they clicked the link). Although this design is not identical to a user receiving no training, the amount of explicit information conveyed to control group users is small.

Because UCSD Health has multiple avenues for users to report potential phishing messages, we did not study whether training improves phishing reporting speeds within the organization. For example, prior work [28] suggests that against mass-phishing emails, user reporting might serve as an effective defense, since a subset of users rapidly reports malicious emails. Future work could explore whether training helps reinforce or improve these reporting dynamics; however, such work should also weigh whether the resulting changes lead to a net improvement for an organization's security, given the potential costs they impose on an organization's IT staff (e.g., responding to user reports and increases in false positive reports [8], [32], [38]).

Finally, although our study and others [14] use link click-through as its metric for phishing failure, user performance may differ based on other metrics, such as whether they would also enter their credentials on the subsequent phishing webpage and/or whether a user would download and execute an attachment.

Future Directions: Assuming our results, and the concurring findings of other work measuring real-world efficacy [7], [28], generalize to most organizations, our work suggests that existing cybersecurity training methods are unlikely to offer practical value towards improving phishing outcomes. Any aspirational attempt to redeem training will need to address several different challenges.

First, existing mechanisms fail to deliver useful training to users. Mandatory annual training offers little impact and embedded training, while in some cases effective, engages only a small fraction of users. Thus, any attempt to improve training must find a way to achieve the uptake of mandatory training, without losing the "teachable moment" experience of embedded training. For example, future work could consider whether adding stronger incentives, either rewards or punitive measures, could improve users' engagement with the training material. However, such research should carefully weigh the ethical and legal implications of this work (e.g., subjecting users to different incentives); and we note that recent studies suggest that making training mandatory or adding positive incentives does not improve the efficacy of training [27]. Furthermore, both anecdotally and based on the results of our mandatory annual training, certain incentives (such as requiring users to complete training) do not always lead to higher intellectual engagement, but may instead incentivize subtle compliance strategies where users simply complete the requirement without actively engaging with the material.

Second, the only modality we identified offering significant improvements in outcomes was interactive training taken to completion. In our study, only a small subset of individuals completed such training. However, it is an open question if a larger cohort could be enticed to similarly engage with such training to completion and if these same effects would generalize to this larger population. If so, this change could produce a modest (19%) but significant improvement for all users. However, it is also entirely possible that the improvement we observed was due to selection effects and that the few users who voluntary completed the interactive training might have have distinct characteristics that make such training more effective for them (e.g., that they are more compliant and are willing to prioritize acquiring computer security knowledge when directed to).

Finally, we cannot disprove the notion that there may be vastly distinct training methods or content that might produce significantly better results (e.g., individualized, real-time training delivered by live instructors who can offer tailored explanations about a user's misunderstandings). However, in addition to the challenge of discovering these new methods, to become practical they must also be economically feasible to deploy at scale.

Absent such advances, our results undercut the fundamental notion that anti-phishing training is a cost-effective endeavor. For an organization with finite resources, it seems likely that focusing on technical countermeasures may offer a better return on investment. In particular, hardware multi-factor authentication (MFA) and/or relying on password managers to fill in credentials only on the correct domain offers strong protections against certain classes of phishing attacks (albeit with a number of operational and usability challenges that deserve further attention [34]).

8. Conclusion

In this work, we analyzed eight-months of simulated phishing campaigns at a large healthcare organization to understand the efficacy of two ubiqituous forms of security training: annual cybersecurity awareness training and embedded phishing training. Through a combination of careful statistical modeling and randomized controlled comparisons, our study finds that both types of training, as commonly deployed today, are unlikely to improve widespread protection against phishing attacks. Although embedded training has a statistical correlation with a lower phishing failure rate, the success rate of many phishing attacks dwarfs the marginal improvement provided by training.

Our paper also investigates understudied aspects of embedded phishing training, such as how much users engage with training content in-the-wild and whether greater engagement with the training material correlates with improved outcomes. In doing so, we shed new light that helps reconcile seemingly contradictory results in prior literature,

and provides guidance on how future studies should structure their experiments and analysis. In particular, future work should focus on using randomized controlled studies, employing training styles that provide greater opportunity for learning (e.g., interactive), and study ways to naturally increase user engagement with the material. Combined with the bulk of empirical evidence from other studies involving real-world, controlled experiments, our results suggests that organizations should not expect large anti-phishing benefits from either annual security awareness training or embedded phishing as commonly deployed today.

Acknowledgements

We thank our anonymous reviewers and shepherd for their time and insightful and constructive feedback. We especially thank UC San Diego and UC San Diego Health's security team and leadership for supporting this research, and Scott Currie, Yvonne Johnson, Phillip Lopo, Daniel Ratcliffe, Minh Vo, Ken Wottge, and Ronise Zenon in particular for their help. This work would not have been possible without their time and effort! Many thanks also to Cindy Moore and Jennifer Folkestad for operational support, and to Enze Liu for help with Latex tooling. This work was supported in part by funding from the University of California Office of the President (UCOP) "Be Smart About Safety" program (an effort focused on identifying best practices for reducing the frequency and severity of systemwide insurance losses), NSF grant CNS-2152644, the UCSD CSE Postdoctoral Fellows program, the Irwin Mark and Joan Klein Jacobs Chair in Information and Computer Science, the CSE Professorship in Internet Privacy and/or Internet Data Security, a generous gift from Google, and operational support from the UCSD Center for Networked Systems.

References

[1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Computing Surveys (CSUR)*, 50(3):1–41, August 2017.

[2] Steve Alder. Security Breaches in Healthcare in 2023. The HIPAA Journal, https://www.hipaajournal.com/security-breaches-in-healthcare/, January 2024.

[3] Sinchul Back and Rob T. Guerette. Cyber Place Management and Crime Prevention: The Effectiveness of Cybersecurity Awareness Training Against Phishing Attacks. *Journal of Contemporary Criminal Justice*, 37(3):427–451, March 2021.

[4] Benjamin Berens, Kate Dimitrova, Mattia Mossano, and Melanie Volkamer. Phishing awareness and education – When to best remind? In *Proceedings of the Symposium on Usable Security and Privacy (USEC)*, pages 1–15, San Diego, CA, USA, April 2022.

[5] Lina Brunken, Annalina Buckmann, Jonas Hielscher, and M. Angela Sasse. "To Do This Properly, You Need More Resources": The Hidden Costs of Introducing Simulated Phishing Campaigns. In *Proceedings of the 32nd USENIX Security Symposium*, pages 4105–4122, Anaheim, CA, USA, August 2023.

[6] Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Benjamin Reinhimer. NoPhish App Evaluation: Lab and Retention Study. In *Proceedings of the Workshop on Usable Security and Privacy (USEC)*, pages 1–10, San Diego, CA, USA, February 2015.

[7] Deanna D. Caputo, Shari Lawrence Pfleeger, Jesse D. Freeman, and M. Eric Johnson. Going Spear Phishing: Exploring Embedded Training and Awareness. *IEEE Security & Privacy*, 12(1):28–38, January/February 2014.

[8] Xiaowei Chen, Sophie Doublet, Anastasia Sergeeva, Gabriele Lenzini, Vincent Koenig, and Verena Distler. What Motivates and Discourages Employees in Phishing Interventions: An Exploration of Expectancy-Value Theory. In *Proceedings of the Twentieth Symposium on Usable Privacy and Security (SOUPS)*, pages 487–506, Philadelphia, PA, USA, August 2024.

[9] Cornell Legal Information Institute. 45 CFR § 164.530 - Administrative requirements. https://www.law.cornell.edu/cfr/text/45/164.530. Accessed 2024-09-26.

[10] Department of Health & Human Services, Healthcare & Public Health Sector Coordinating Council. Technical Volume 2: Cybersecurity Practices for Medium and Large Healthcare Organizations. https://405d.hhs.gov/Documents/tech-vol2-508.pdf, 2023 Edition.

[11] Department of Health & Human Services, Office of Information Security. Ransomware and Healthcare. https://www.hhs.gov/sites/default/files/ransomware-healthcare.pdf, January 2024.

[12] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why Phishing Works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 581–590, Montréal, Québec, Canada, April 2006.

[13] Executive Office of the President of the United States. Federal Information Security Modernization Act of 2014. Annual Report to Congress, Fiscal Year 2021, https://www.whitehouse.gov/wp-content/uploads/2022/09/FY2021-FISMA-Report-to-Congress.pdf, September 2022.

[14] Anjuli Franz, Verena Zimmermann, Gregor Albrecht, Katrin Hartwig, Christian Reuter, Alexander Benlian, and Joachim Vogt. SoK: Still Plenty of Phish in the Sea—A Taxonomy of User-Oriented Phishing Interventions and Avenues for Future Research. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security (SOUPS)*, pages 339–357, August 2021.

[15] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, December 2006.

[16] William J. Gordon, Adam Wright, Robert J. Glynn, Jigar Kadakia, Christina Mazzone, Elizabeth Leinbach, and Adam Landman. Evaluation of a mandatory phishing training program for high-risk employees at a US healthcare system. *Journal of the American Medical Informatics Association (AMIA)*, 26(6):547–552, June 2019.

[17] Cormac Herley. So Long, and No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. In *Proceedings of the New Security Paradigms Workshop (NSPW)*, pages 133–144, Oxford, United Kingdom, September 2009.

[18] Doron Hillman, Yaniv Harel, and Eran Toch. Evaluating organizational phishing awareness training on an enterprise scale. *Computers & Security*, 132, 2023.

[19] David W. Hosmer Jr., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, Third edition, March 2013.

[20] IBM Security. Cost of a Data Breach Report 2023. https://www.ibm.com/downloads/cas/E3G5JMBP, July 2023.

[21] Daniel Jampen, Gürkan Gür, Thomas Sutter, and Bernhard Tellenbach. Don't click: towards an effective anti-phishing training. A comparative literature review. *Human-centric Computing and Information Sciences*, 10(33):1–41, August 2020.

[22] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. School of Phish: A Real-World Evaluation of Anti-Phishing Training. In *Proceedings of the 5th Symposium on Usable Privacy and Security (SOUPS)*, pages 1–12, Mountain View, CA, USA, July 2009.

[23] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Protecting People from Phishing: The Design and Evaluation of an Embedded Training Email System. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 905–914, San Jose, CA, USA, April 2007.

[24] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Lessons From a Real World Evaluation of Anti-Phishing Training. In *APWG Annual eCrime Researchers Summit (eCrime)*, pages 1–14, Atlanta, Georgia, USA, October 2008.

[25] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching Johnny Not to Fall for Phish. *ACM Transactions on Internet Technology (TOIT)*, 10(2):1–31, June 2010.

[26] Eunkyung Kweon, Hansol Lee, Sangmi Chai, and Kyeongwon Yoo. The Utility of Information Security Training and Education on Cybersecurity Incidents: An empirical evidence. *Information Systems Frontiers*, 23:361–373, April 2021.

[27] Daniele Lain, Tarek Jost, Sinisa Matetic, Kari Kostiainen, and Srdjan Capkun. Content, Nudges and Incentives: A Study on the Effectiveness and Perception of Embedded Phishing Training. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, pages 1–17, Salt Lake City, UT, USA, October 2024.

[28] Daniele Lain, Kari Kostiainen, and Srdjan Čapkun. Phishing in Organizations: Findings from a Large-Scale and Long-Term Study. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 842–859, San Francisco, CA, USA, May 2022.

[29] Marsh McLennan. Using data to prioritize cybersecurity investments. https://www.marsh.com/en/services/cyber-risk/insights/using-cybersecurity-analytics-to-prioritize-cybersecurity-investments.html, 2023.

[30] Nina Marshall, Daniel Sturman, and Jaime C. Auton. Exploring the evidence for email phishing training: A scoping review. *Computers & Security*, 139:103695:1–16, April 2024.

[31] National Institute of Standards and Technology. Security and Privacy Controls for Information Systems and Organizations. NIST 800-53 Rev. 5.1.1. https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final, November 2023.

[32] Nikolas Pilavakis, Adam Jenkins, Nadin Kokciyan, and Kami E Vaniea. "I didn't click": What users say when reporting phishing. In *Proceedings of the Symposium on Usable Security and Privacy (USEC)*, pages 1–13, San Diego, CA, USA, February 2023.

[33] Posit. *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA, USA, 2024.

[34] Ken Reese, Trevor Smith, Jonathan Dutson, Jonathan Armknecht, Jacob Cameron, and Kent Seamons. A Usability Study of Five Two-Factor Authentication Methods. In *Proceedings of the Fifteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 357–370, Santa Clara, CA, USA, August 2019.

[35] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana Von Landesberger, and Melanie Volkamer. An investigation of phishing awareness and education over time: When and how to best remind users. In *Proceedings of the Sixteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 259–284, August 2020.

[36] Orvila Sarker, Sherif Haggag, Asangi Jayatilaka, and Chelsea Liu. Personalized Guidelines for Design, Implementation and Evaluation of Anti-Phishing Interventions. In *Proceedings of the International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–12, New Orleans, LA, USA, October 2023.

[37] Orvila Sarker, Asangi Jayatilaka, Sherif Haggag, Chelsea Liu, and M. Ali Babar. A Multi-vocal Literature Review on challenges and critical success factors of phishing education, training and awareness. *Journal of Systems & Software*, 208:111899:1–25, March 2024.

[38] Katharina Schiller, Florian Adamsky, Christian Eichenmüller, Matthias Reimert, and Zinaida Benenson. Employees' Attitudes towards Phishing Simulations: "It's like when a child reaches onto the hot hob". In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, pages 1–15, Salt Lake City, UT, USA, October 2024.

[39] Markus Schöps, Marco Gutfleisch, Eric Wolter, and M. Angela Sasse. Simulated Stress: A Case Study of the Effects of a Simulated Phishing Campaign on Employees' Perception, Stress and Self-Efficacy. In *Proceedings of the 33rd USENIX Security Symposium*, pages 4589–4606, Philadelphia, PA, USA, August 2024.

[40] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish. In *Proceedings of the Third Symposium on Usable Privacy and Security (SOUPS)*, pages 88–99, Pittsburgh, PA, USA, July 2007.

[41] Hossein Siadati, Sean Palka, Avi Siegel, and Damon McCoy. Measuring the Effectiveness of Embedded Phishing Exercises. In *Proceedings of the 10th USENIX Workshop on Cyber Security Experimentation and Test (CSET)*, pages 1–8, Vancouver, BC, Canada, August 2017.

[42] Kai Florian Tschakert and Sudsanguan Ngamsuriyaroj. Effectiveness of and user preferences for security awareness training methodologies. *Heliyon*, 5(6):e02010:1–10, June 2019.

[43] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, Philipp Rack, Marco Ghiglieri, Peter Mayer, Alexandra Kunz, and Nina Gerber. Developing and Evaluating a Five Minute Phishing Awareness Video. In *Proceedings of the 15th International Conference on Trust, Privacy and Security in Digital Business (TrustBus)*, pages 119–134, Regensburg, Germany, September 2018. Springer.

[44] Melanie Volkamer, Martina Angela Sasse, and Franziska Boehm. Analysing Simulated Phishing Campaigns for Staff. In *Proceedings of the 2nd Workshop on Security, Privacy, Organizations, and Systems Engineering*, pages 312–328, Guildford, UK, September 2020.

[45] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 108:1–12, Glasgow, Scotland, UK, May 2019.

[46] Tianjian Zhang. Knowledge Expiration in Security Awareness Training. In *Proceedings of the Annual ADFSL Conference on Digital Forensics, Security and Law (CDFSL)*, pages 197–212, San Antonio, TX, USA, May 2018.

Appendix A.

Phishing Messages and Training Programs

Phishing Messages: For our phishing campaigns, we used templates from Proofpoint's Drive-By campaign library. We provide a brief summary of message in each lure below, along with the exact name of the template in Proofpoint's library (as of October 2024):

1) Outlook Pwd: This campaign purports to come from the IT Security team and states that the user's account has been suspended due to suspicious activity. The user must click on a link to reset their account. (Proofpoint Title: "Outlook Account Reset")

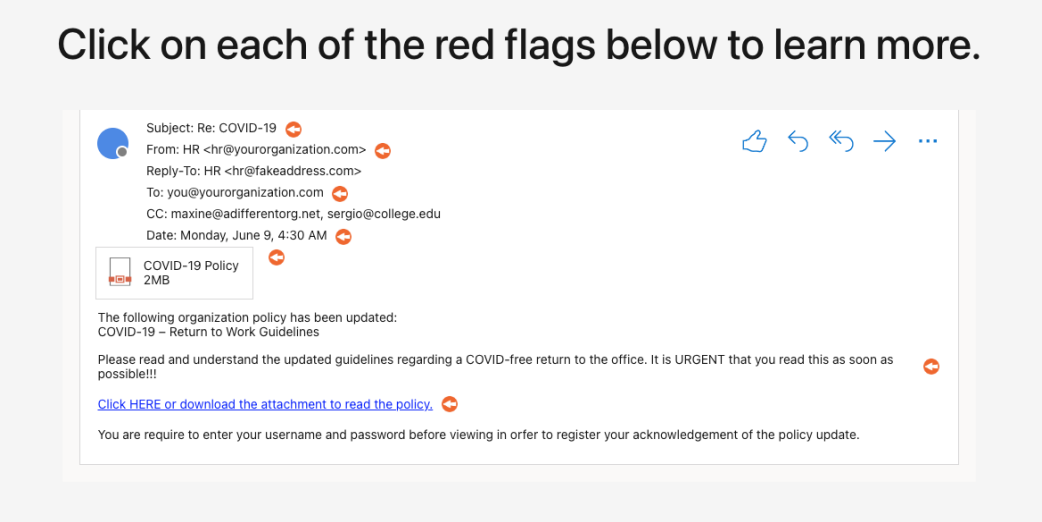


Figure 5: An example exercise included in the *annual cybersecurity awareness* training (§ 4).

2) Login Account: This campaign appears to come from IT Support. It alerts the user that the IT department has discovered their password was stolen by hackers. The user must click on a link to reset their password to avoid having their access revoked. (Proofpoint Title: "Login Account Compromised")
3) Open Enroll: This campaign purports to comes from Human Resources. It notifies users that the Open Enrollment for benefits is approaching and provides a number of benefit related links for the user to click. (Proofpoint Title: "Open Enrollment Update")
4) Shared Doc (Microsoft): This campaign comes from a random sender and notifies the recipient that the sender has shared a document with them via OneDrive. The user can click on a link to "view" the document. (Proofpoint Title: "OneDrive Document Waiting")
5) OneDrive Medical: This campaign comes from a random sender with a "Dr." prefix. The email has OneDrive logos and notifies the user that a file has been shared and needs to be reviewed asap. (Proofpoint Title: "OneDrive Medical Document")
6) Docusign: This campaign comes from a random sender, with a domain that sounds like a financial institution. The email body mimics a DocuSign email with a link to review the shared documents. (Proofpoint Title: "Docusign document for review")
7) Building Evac: This campaign comes from the HR Department and notifies the user that there is an updated building evacuation plan. It provides a link for the users to view the plan and states that the user must digitally sign an acknowledgement of the new plan. (Proofpoint Title: "Building Evacuation Plan")
8) Traffic Ticket: This campaign comes from a generic traffic enforcement entity. It states that the user has a speeding ticket and must click a link to view and pay the fine to prevent it from increasing. (Proofpoint Title: "Speeding Violation")
9) Dress Code: This campaign comes from a random sender from the human resources department. It says that a new dress code policy will go into effect and provides a link to view the new policy, along with a warning of disciplinary action for violations in the future. (Proofpoint Title: "Dress Code")
10) Vacation Policy: This campaign comes from human resources and notifies the recipient that there is a new vacation and sick time policy. (Proofpoint Title: "Updated vacation and sick time policy")

Phishing Training: For the study's four training programs, we used two stock trainings from Proofpoint's "Teachable Moment" library of embedded training materials: the "Standard: Education (Drive-by)" and "Interactive URL Guide" for our generic static and interactive trainings respectively.

The static training stated that the user fell for a phishing simulation, that the attacker tried to make the user click on a malicious link, provided four tips for avoiding future attacks, and included a link at the bottom of the webpage to "acknowledge" and close the training. The four tips asked users to check for grammar/spelling errors, check the email language for an unusual tone, hover over any links, and use a "reputable source" to "verify the link".

The interactive training also notified users that they fell for a phishing simulation, and then launched users into an interactive series of training webpages. The webpage displayed the same sample phishing email on each page, but asked the users to a new question on each page that corresponded to a phishing warning sign (such as typos, lack of a personalized greeting, and whether the email contained a dangerous action like a link to click).

For our two contextualized / customized training programs, we modified these two trainings to include a consistent set of five tips to spot phishing attacks, selected from a list of advice from industry and government websites: checking whether the email comes from an unusual sender,

contains a dangerous action (such as clicking a link or attachment), includes threats or urgent language, uses a generic greeting without any personalization/naming of the user, and/or contains grammar or spelling errors.

In the contextual static training, we modified Proofpoint's training template to include an image of the phishing email the recipient received. Our training then highlighted places in the email where a piece of advice applied (one of the five warning signs), and included a 2-3 line summary of which warning signs were present above the "acknowledge" button.

For the contextual interactive training, we replaced the sample phishing email with the actual phishing email a user received. We then modified the question on each slide to correspond to ask about the presence of one of the five warning signs, and to grade and display the correct answer accordingly.

Appendix B.
Additional Figures

Figure 5 shows an example of one part of the annual cybersecurity awareness training deployed at UCSD Health.

Appendix C.
Meta-Review

The following meta-review was prepared by the program committee for the 2025 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

C.1. Summary

This paper evaluates the effectiveness of phishing simulations using an in-situ randomized controlled experiment with 19,500 employees in an organization over eight months. It was found that embedded training had no significant effect on the security behavior of the trained employees.

C.2. Scientific Contributions

- Independent Confirmation of Important Results with Limited Prior Research
- Provides a Valuable Step Forward in an Established Field
- Addresses a Long-Known Issue

C.3. Reasons for Acceptance

1) This work presents the most extensive large-scale evaluation of phishing simulations to date. It confirms a previous study by Lain et al. that cast doubt on the effect of phishing simulations against common belief.
2) The authors were also the first to perform a well-designed, large-scale, randomized, controlled experiment about phishing simulations in an organization. This method has multiple advantages over lab-like studies previously used to analyze anti-phishing user behavior, as it gatherers the data in the real-world context of the users (employees).
3) The paper has a real-world impact regarding cybersecurity training for employees. It questions the "best-practice" of embedded training in phishing simulations that are rolled out by multiple organizations.

C.4. Noteworthy Concerns

1) The authors cannot provide the raw data (due to the organization's policies). Hence, independent confirmation of the statistical tests is not possible.
2) As with every in situ study, some findings might be tied to the specific context of the organization they are carried out at (a health company in this case).