Phishing for Significance

Iris Lew, Angelique Iradukunda, Srila Maiti

12/01/2022

Abstract

Is active learning more effective than passive learning in terms of educating college students against phishing emails offering employment opportunities? College students are key targets to phishing scams because it is often their first time searching for employment and many expect to find jobs upon graduation. Thus, we seek to improve the quality of cybersecurity training materials to educate students on identifying malicious emails. We test whether the type of learning (passive vs. active education) will successfully enable students to identify whether emails are phishing or legitimate.

We distributed a survey to a sample marketplace, Pure Spectrum. Survey takers will either have suspicious parts in emails identified for them (passive) or have to click the suspicious parts of an email (active). Afterwards, they will view a series of 10 emails where they have to identify each as legitimate or phishing. We find no difference between the accuracy rates of the two groups.

Background

As reported by many surveys, phishing has been around since 1995; it started off by phishers stealing user's passwords and using algorithms to generate fraudulent accounts through fake credit card numbers[1]. Generally, it has been observed that there is an increase of misleading emails and websites as individuals become more connected to the virtual world; it is not promising that phishing is going to end.

Digital scams have evolved and students are common targets. While many companies include phishing awareness training as part of their cybersecurity training, students often do not receive such training despite the fact that most students will likely end up looking for job opportunities when they graduate and thus are particularly vulnerable to online recruitment fraud (ORF). There are employment scams, student loans relief scams, fraudulent scholarship opportunities, and many more which are specifically aimed at them[2]. Because of this, colleges and universities are aware that they need to work on security solutions to help improve students' cybersecurity knowledge, including phishing awareness. Previous research has been done to show that simulated phishing emails to Carnegie Mellon University students, staff and faculty as of the training will be beneficial, but it was done through a program called PhishGuru[3]. Currently, the university we are attending, UC Berkeley, carries a repository of previous phishing emails and a few posters, but does not require any training. Students can read materials on how to "Fight the Phish"[4], but usually the only people who read materials are those who are already interested in the first place. Additionally, it is not just students who are interested in cybersecurity who should learn about phishing, as students in general are receiving phishing emails.

AT&T states that there are two parts to phishing awareness training: awareness education and phishing testing[5]. We are interested in whether reading materials about cybersecurity would be just as effective,

[1] https://www.phishing.org/history-of-phishing
[2] https://www.fdacs.gov/Consumer-Resources/Scams-and-Fraud/Scams-Targeting-College-Students
[3] https://www.cs.cmu.edu/ jasonh/publications/acm-tois-teaching-johnny-not-to-fall-for-phish-final.pdf
[4] https://security.berkeley.edu/education-awareness/phishing/fight-phish-marketing-materials
[5] https://cybersecurity.att.com/blogs/security-essentials/phishing-awareness-training-explained

or whether it would be more effective if students had to interact with them. Research has been done to show active learning ("the process of having students engage in some activity that forces them to reflect upon ideas and how they are using those ideas") improves learning compared to passive methods[6]. Previous research also shows that while incorporating screenshots of potential phishing emails might not help train people on categorizing phishing emails and legitimate emails even though it increases awareness[7], using a game called Anti-Phishing Phil would help people distinguish legitimate websites from phishing websites[8], suggesting that interactivity with the training materials might help. Thus, in this project, we are studying whether this improvement would extend to teaching students, prime targets, about phishing in the context of cybersecurity.

Research Question

As students who also have a professional background, we are wary when we are solicited by recruiters and receive job offers because we also receive phishing emails. While there is technology that helps to filter out phishing emails[9], not all are filtered out and humans are still capable of making errors due to ignorance. If we become too reliant on this technology, we could potentially fall prey to the few phishing scams that go through.

So we seek to improve on current cybersecurity phishing materials. Is an active learning approach where students have to interact with training materials more effective? Our experiment aims to investigate the causal effect of active or passive learning methods for cybersecurity training to promote phishing awareness among students. If we use a learning technique that challenges students to retain information better, would it help improve students' ability to identify phishing emails? We operationalize "identifying phishing emails" as the percentage of emails correctly identified out of 10 emails (5 of which are recruiting or job offer emails that we have received, 5 of which are phishing emails).

Hypothesis

In order to understand the effectiveness of cybersecurity training for identifying emails as phishing or legitimate correctly, we start by establishing our below hypotheses:

- **Null Hypothesis (H_0)**: There is no statistically significant difference in rate of accurately identifying emails as legitimate or phishing between students who are receiving active or passive training.
- **Alternate Hypothesis (H_a)**: There is a statistically significant difference in the rate of accurately identifying the emails as legitimate or phishing between students who are receiving active or passive training.

Based on the background research, we expect the students going through interactive cybersecurity training will be able to identify phishing emails more than those receiving passive training, but we are being conservative and only testing for whether a difference can be found.

Experiment Design

Experiment Overview

To test this hypothesis, we sent a survey and used a between-subjects design. The survey collected various demographic attributes which we think could be used as covariates. Participants will be randomly assigned to either the treatment (active) or control (passive) and will go through their appropriate training procedures.

We combined training type (active or passive) and whether a timer is used on a particular training topic in a 2x2 factorial design. We wanted to discern whether it is the time spent in the active training which would

[6]https://journals.physiology.org/doi/full/10.1152/advan.00053.2006
[7]https://link-springer-com.libproxy.berkeley.edu/content/pdf/10.1007/978-3-540-77366-5_33
[8]https://www.cs.cmu.edu/ jasonh/publications/acm-tois-teaching-johnny-not-to-fall-for-phish-final.pdf
[9]https://workspace.google.com/blog/identity-and-security/an-overview-of-gmails-spam-filters

increase the accuracy (because respondents would have to read the training carefully and search for where to click), or whether it is because the interaction itself helps them retain the information better. Participants who did not have a timer could progress at their own pace but those with a timer had to spend at least 20 seconds on the pages. This leads us the design below:

Passive Training X No Timer	Passive Training X Timer Imposed (20 secs)
Active Training X No Timer	Active Training X Timer Imposed (20 secs)

Upon completion of the training, the participants will then be presented a series of emails where they have to determine if they are legitimate or phishing emails. Based on their responses, we will evaluate whether our hypothesis holds true or not.

Comparison of Potential Outcome

To control for both measured and unmeasured potential confounding variables, we randomly assigned students in the treatment (active learning) group or control (passive learning) group. We consider the two different types of treatments for our participants: the first being active treatment and the second being passive treatment; we then compare the outcome of the percentage of emails identified correctly between the two groups. The potential outcome for treatment is the percentage of correctly identified emails with active conditions as opposed to potential outcome for control which is the percentage of emails correctly identified with passive condition.

We hope to find a difference in the potential outcomes for the treatment compared to the potential outcomes for the control. In other words, we hope that the average percentage of emails that are correctly identified would be different between those in the active learning group compared to those in the passive learning group.

Survey Instrument - Qualtrics

We built the survey instrument in Qualtrics and divided it into four main parts: the demographics including metadata, attention check, the training, and the measurement.

The demographics would include screener questions in order to ensure we only keep the sample from our target population: current students. We decided to ask questions about the students' residential status and if they were actively seeking jobs or internships in order to include them as covariates. We chose these covariates because we suspect that certain groups of these students may make them better targets for phishers (e.g., an international student may not be as familiar with job hunting in the US).

The attention check required respondents to read and answer a math question. Two numbers between 0 and 9 were generated. The larger number would be the number of balls that were found on a floor in the question while the smaller number functioned as the number of them that a child would pick up. Survey takers were asked how many balls were left on the floor. The training process included a short explanation at the top followed by a screenshot of an email with suspicious elements. If respondents were in the active condition, they would have to click the suspicious elements in the screenshot, which would then be highlighted. If respondents were in the passive condition, they would see a pre-drawn box around the suspicious elements in the screenshot. Furthermore, if respondents were in a timer condition, then they would only be able to proceed after 20 seconds on each page. There were four pages to the training process, and they would be presented in a random order to the participants.

The measurement emails were recreated so that they appeared more standardized, sent to our own email inboxes, and then included in Qualtrics as screenshots. The legitimate emails were sourced from our own email inboxes and LinkedIn solicitations from recruiters and hiring managers. The phishing emails were sourced from the UC Berkeley "Phish Tank" [10] and the Georgetown University Information Security Office

[10] https://security.berkeley.edu/resources/phish-tank

Phishing Examples[11]. We started with 10 legitimate emails and 10 phishing emails, but after considering respondent fatigue, we decided to limit it to 5 legitimate emails and 5 phishing emails. These 10 emails were also presented in a randomized order. Furthermore, the order of the answer choices of "legitimate" and "phishing" were randomized to control for possible order effects. We used Qualtrics built-in scoring system to identify how many emails are correctly marked and then calculated the percentage correct within Qualtrics for each respondent.

Randomization

Once a survey taker passes the attention check, they would be randomly assigned to passive or active learning using Qualtrics's built-in randomization function within its survey flow. Then they would be assigned to a timer condition using the same randomization function, where they can progress at their own pace or they have to spend 20 seconds at each of the training pages before they could proceed. This was the only point when they could be assigned into these conditions, and because the randomization happened without the respondents' knowledge and respondents could only proceed forwards in the survey, there was no noncompliance. This provided a 2x2 multi-factorial experiment with the following cell sizes that are used for quotas.

	Progressed At Own Pace	Mandatory Timer (20 Sec)
Passive	Pilot n=13; Full n=65	Pilot n=13; Full n=65
Active	Pilot n=11; Full n=65	Pilot n=13; Full n=66

We aimed to have 50 completed surveys in the pilot and 260 in the full, following the power analysis. Quotas allowed 13 for each group in the pilot, and 65 for each group in the full. We ended up with 66 completes in the Active + Timer condition because two people submitted a response at around the same time, thus incrementing the quota together.

While the assignment is initially random, as quotas filled up, we would redirect them to other quotas. For example, if someone were assigned to Passive + No Timer, but Passive + No Timer and Passive + Timer had their quotas filled, they would then be funneled over to Active + No Timer. If the quota were met and they could not be funneled over, then they were redirected out of the survey. We instituted this redirect because we believed survey takers were independent of one another and the next respondent that would appear in our sample set would be required to be the other condition.

Recruitment Process

We decided to use Pure Spectrum as a supplier platform. There are a total of 50 suppliers who could send survey takers to our survey. Pure Spectrum has pre-built screeners, and we decided to pass in respondents' answers to the age, gender, and employment status screeners into our data. Pure Spectrum's screener asked "What is your employment status?" and one of the response options is "Student." We chose not to use their employment status as a criterion to exclude respondents because it is a single-select and we were interested in including all students, including part-time students who are also full-time workers. In our survey tool, we included two screener questions. The first screener asked the age of the respondent. We decided to exclude respondents who report that they are "Under 18" years of age because people have to be over 18 years old in order to be eligible for most jobs in the United States. We have also decided to only include respondents who report that they are an "actively enrolled student." Thus, they would be more likely to have a student email that they regularly check, and are the ones most likely to fall prey to student recruitment scams. Respondents who fail to answer pass these screeners will be redirected out of the survey and would not increase our quota counts.

We also included an attention check after the demographics. If respondents answered incorrectly, they would be redirected out of the survey and would not increase our quota counts. The survey had built in

[11] https://security.georgetown.edu/category/phishing-examples/

randomization as soon as respondents passed the attention check which would increment their respective quotas. The quota would only be incremented if the respondents complete the survey.

We fielded a pilot (n=50) and our full sample (n=261). Individuals who have participated in the pilot, regardless if they were screened out, redirected out due to reaching quota, failed the attention check, or did not complete the survey, would be ineligible to be a part of the full sample. Individuals who completed the survey were compensated with $0.60-$0.70 in the pilot and with $0.70-$1.50 in the full sample.

Project Timeline

Survey Design Finished	Pilot Fielding	Pilot Data Analysis Starts	Full Fielding	Full Analysis Starts
Nov 18, 2022	Nov 19-20, 2022	Nov 20, 2022	Nov 21-24, 2022	Nov 24, 2022

We decided to field a pilot sample of 50 respondents so that we can do a power analysis for the full fielding and account for any unforeseen difficulties in actual fielding.

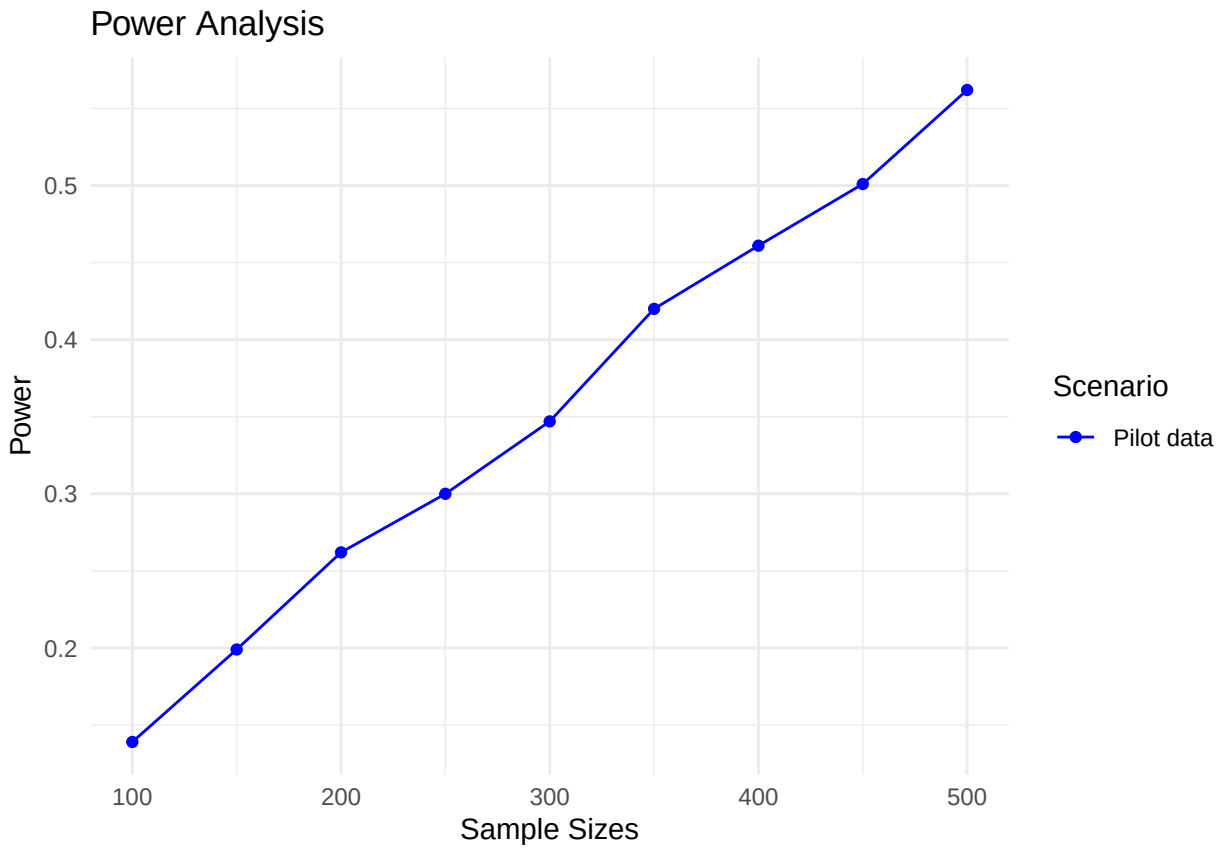
Pilot Study

In the pilot study, we discovered that the plan to redirect respondents to the other condition once quotas were filling up was not implemented. It was discovered towards the end of the study, and updated the survey instrument to include the logic. Otherwise, our questions and measurement calculations were working as intended.

We also used the pilot study to test how we should price the cost per interview and how fast we would get responses. There were more complete responses during non-working hours and when people would typically be awake in the United States; however, it took two days to gather 50 completes, and thus we discovered we needed to provide higher prices and increase it more frequently to attract more sample members.

Power Analysis

We could not find any background research on passive vs. active learning in the context of phishing emails so our power analysis is based on our own pilot study of n=50. We did not find any statistical significance with an n=50 (p-value of 0.49). During our power analysis, we discovered that we would not reach 80% power even if we paid everyone $1. Our financial analysis indicated that we would likely need to pay some respondents more than $1 in order to meet fielding deadlines but with a budget limit of $500, so we decided to field a full sample of 260 respondents using a conservative estimate of our budget.



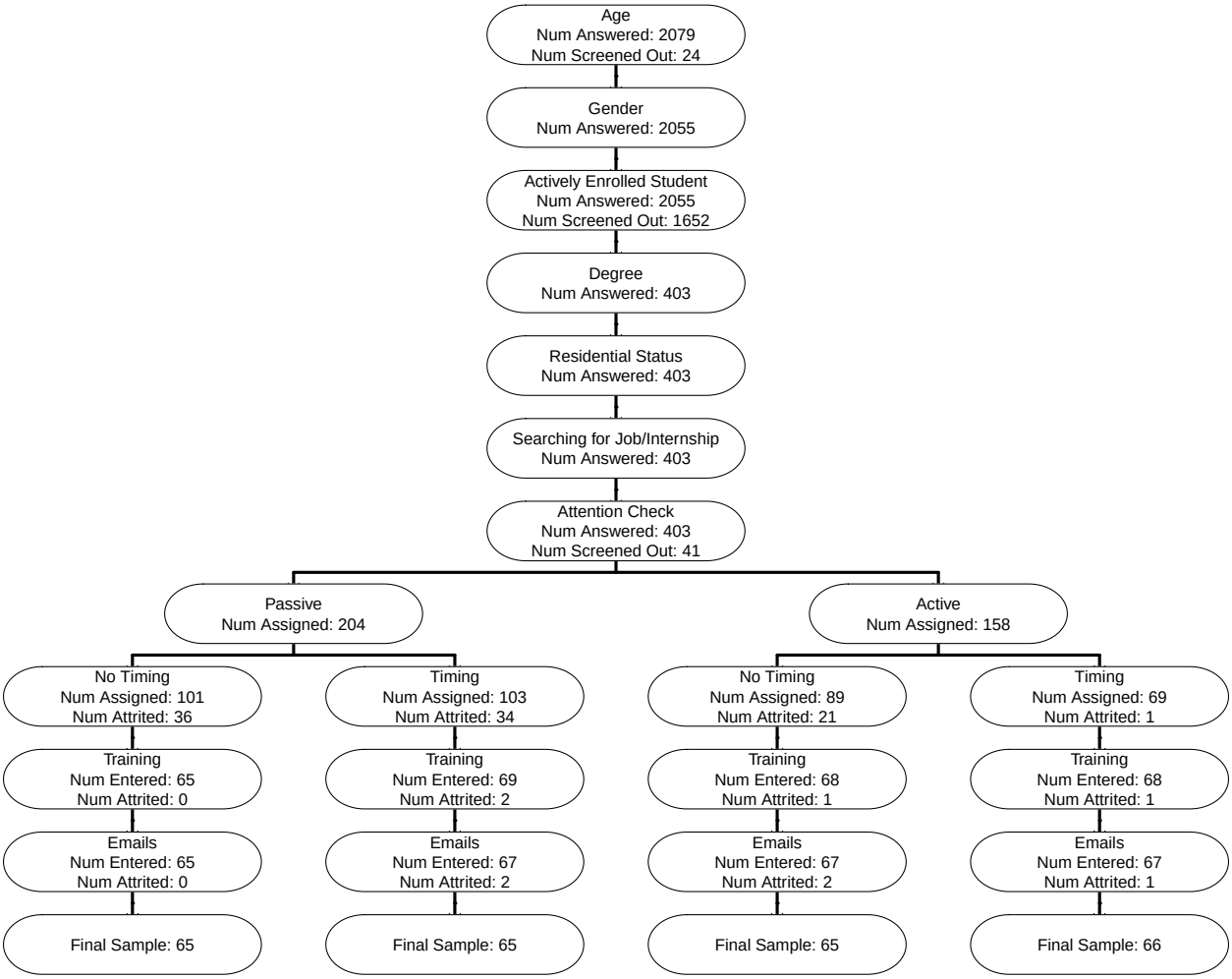
Observation and Outcome Measures

Each observation in our data set is corresponding to one survey taker's response and each survey taker can take the survey only once. Pure Spectrum's platform contains a report which also includes additional attributes which are not part of our survey design, however we have incorporated the data in the final dataset. Respondents did not have to answer every question that is included in Pure Spectrum's screeners, so there is missing information there; however, they had to answer all of the questions in our Qualtrics questionnaire.

Each observation consists of metadata information (browser, OS version, IP address, start and end time, finished or not, duration, latitude, longitude, language etc), demographic information (age, gender, ethnicity, relationship, occupation, highest education, kids, geographic location attributes, currently enrolled students or not, international/in-state/out-of-state student etc), subject assignment information (active/passive X timer/no timer condition), how much time the subject is spending on the training topic/question, what the survey taker classified the emails in the measurement section as, and measurement calculations. There were 10 emails in the measurement section, and the number the respondents answered with the correct response was converted into a percentage.

We had incomplete observations. If they were marked as incomplete before they were assigned to a learning group and a timer condition, we could not count them as part of our sample. If they were assigned a learning group and a timer condition, we would consider them as attrited.

Data Completeness

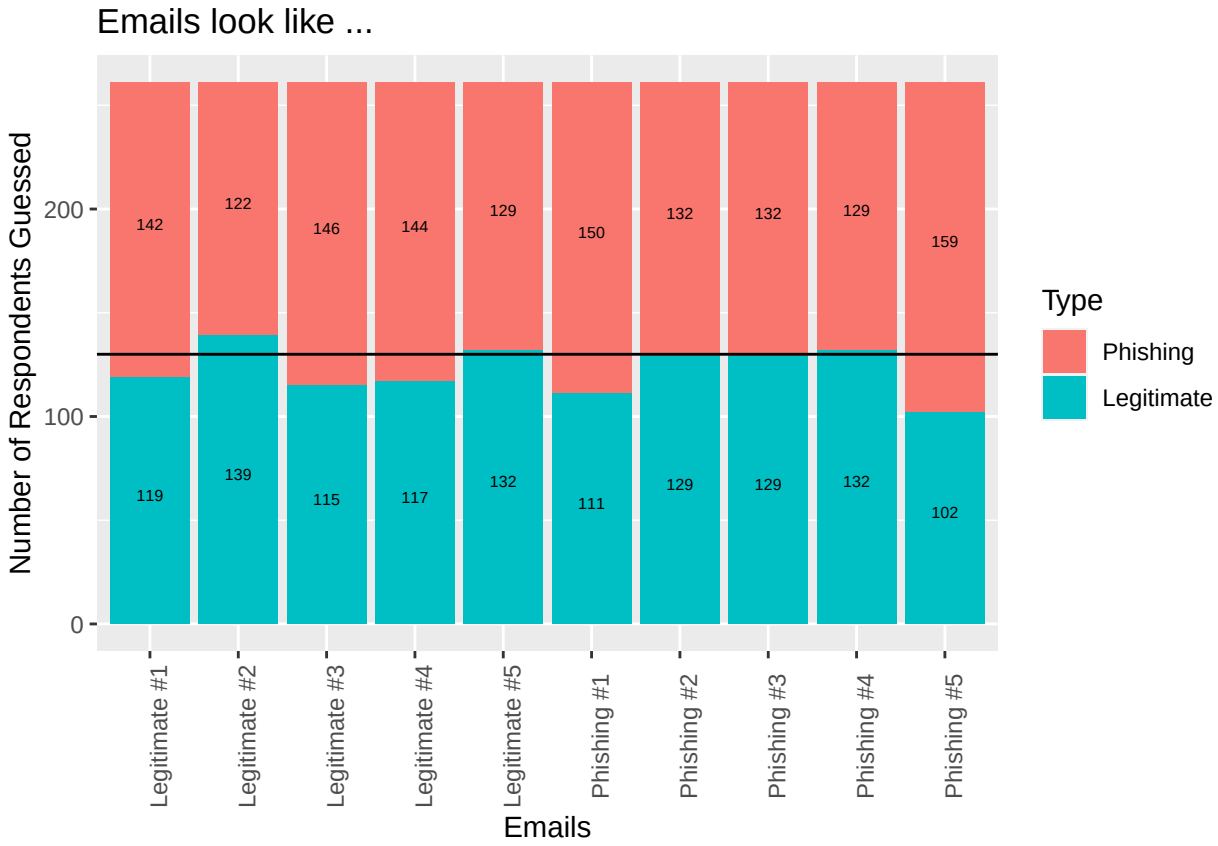


Of the 2079 people who entered the survey, only 362 passed the screeners and were eligible to be a part of the experiment. As mentioned above, because we assigned respondents to their training and timer conditions at one point and they could only move forward in the survey, there was no noncompliance.

We noticed that there were 92 respondents who chose to attrit as soon as they were assigned into a condition, and never received the training at all. Furthermore, there was differential attrition with only 1 person that was assigned into the active and timer condition leaving the survey compared to a combined 91 from the other conditions. The remaining 9 attrited during the training or when they were classifying emails. As there was more than 5% attrition, we did not apply an extreme value bounds analysis. See the diagram above for more details on how many made it through the survey and how many left at which time.

Results

For our analysis, we only used the complete data. First of all, we also took a look at whether emails were identified correctly across the groups to see if there is a trend in phishing emails looking "phishy" and legitimate emails appearing as legitimate.



Overall, the 10 emails performed similarly to each other and the emails tended to look "phishy" in general, but using a chi-square test, we did not find any significant difference (p-value of 0.44, df=18) between the emails and how "phishy" they looked. Thus, the emails are comparable to each other.

We ran three linear regression models:

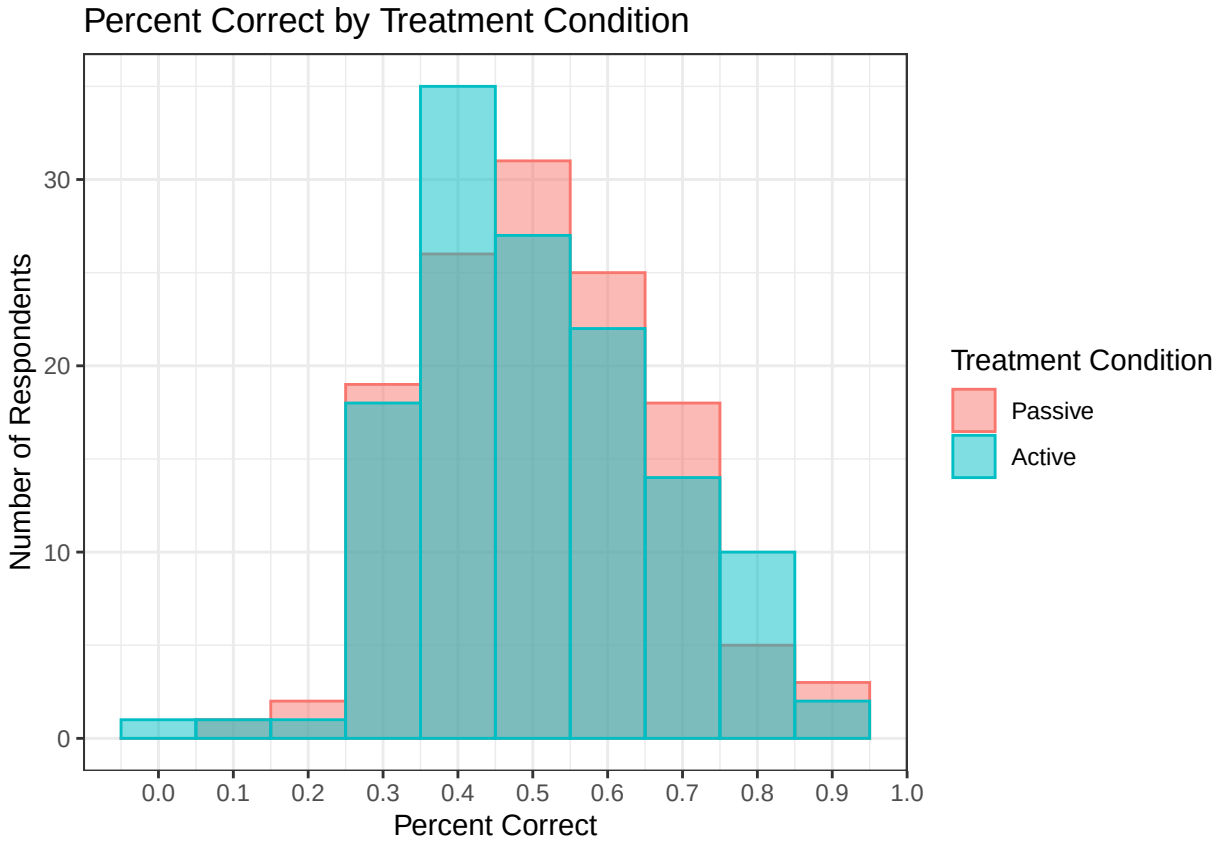
- The effects of the type of training on the percentage correct, with whether a timer was assigned as a covariate.
- The effects of the type of training, interacted with whether a timer was assigned, on the percentage correct. This is so that we can explore possible heterogeneous effects.
- Any possible causal effects from the type of training, whether a timer was assigned, and other features we collected as covariates.

Table 4: Linear Regression Models

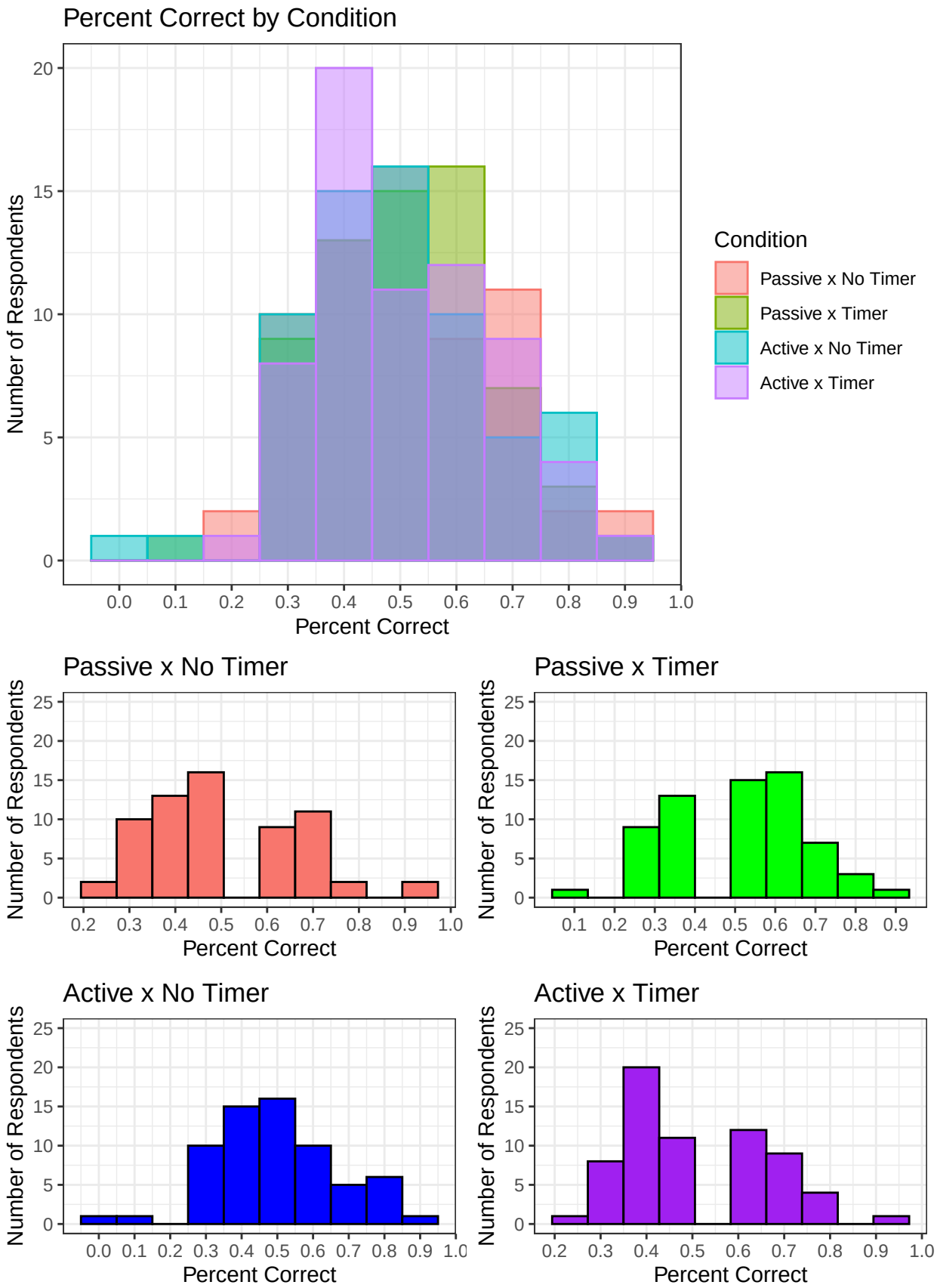
	Percent of Emails Classified Correctly		
	Treatment Design	Treatment Design with Interaction	Exploratory Model
	(1)	(2)	(3)
Active	−0.01	−0.01	−0.03
	(0.02)	(0.03)	(0.03)
Timer	0.01	0.003	−0.004
	(0.02)	(0.03)	(0.03)
Degree: Associate			0.10***
			(0.03)
Degree: Bachelor's			0.13***
			(0.03)
Degree: Master's			0.08**
			(0.04)
Degree: Professional			0.09*
			(0.05)
Degree: PhD			0.05
			(0.06)
Degree: Technical Certifications			0.15**
			(0.07)
Active/Timer Interaction		0.01	0.03
		(0.04)	(0.04)
Constant	0.51***	0.51***	0.37***
	(0.02)	(0.02)	(0.06)
Graduating			✓
Age			✓
Gender			✓
Residential Status			✓
Device			✓
additional features			✓
Observations	261	261	261
R^2	0.001	0.001	0.08
Residual Std. Error	0.16 (df = 258)	0.16 (df = 257)	0.16 (df = 243)

Note: *p<0.1; **p<0.05; ***p<0.01

In our regressions, we see that the results for the active and passive learning conditions are not significant. The average percent correct for those in the passive condition is 0.51 and 0.5 for those in the active condition. We ran the linear regression and did not find any significance. This did not improve by much even when we included the interaction. With robust standard errors and without covariates, we find a p-value of 0.68 with an average treatment effect of -0.0069524 and a standard error of 0.03. In the practical sense, this means that reading about what to look for in a phishing email and interacting with the email as part of training would provide similar amounts of information. This is reflected in the histograms of the data, where we see a similar shape no matter the condition. There is a slight left skew in the data, but we do not see a ceiling effect where everyone is correctly identifying all the emails.

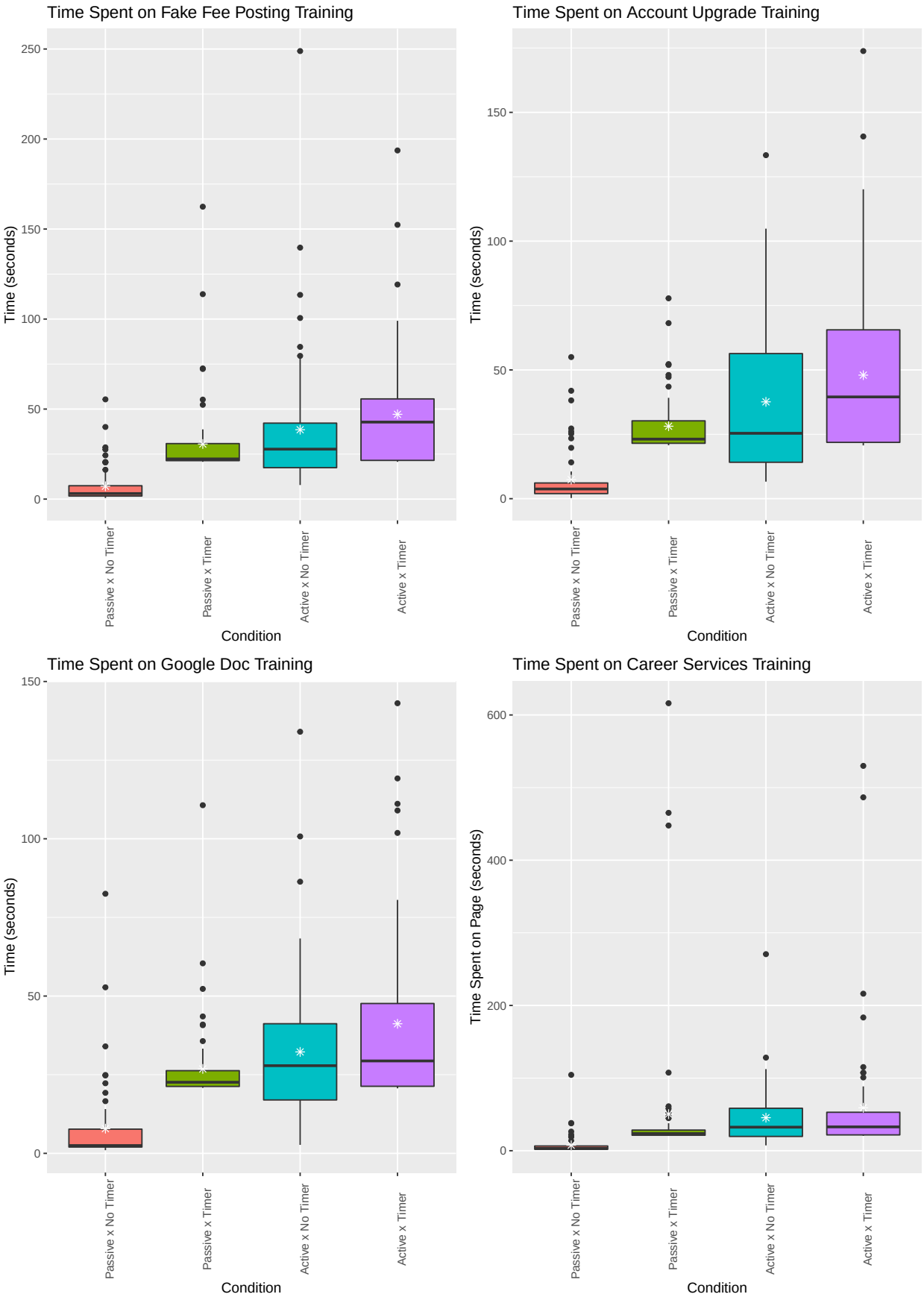


As we have included a timer condition, we can further parse out the data into four conditions. We see that in all four conditions, the histogram of each group's accuracy takes on a similar shape, especially when we layer them on top of each other. We separated the four histograms out in order to better depict the counts.

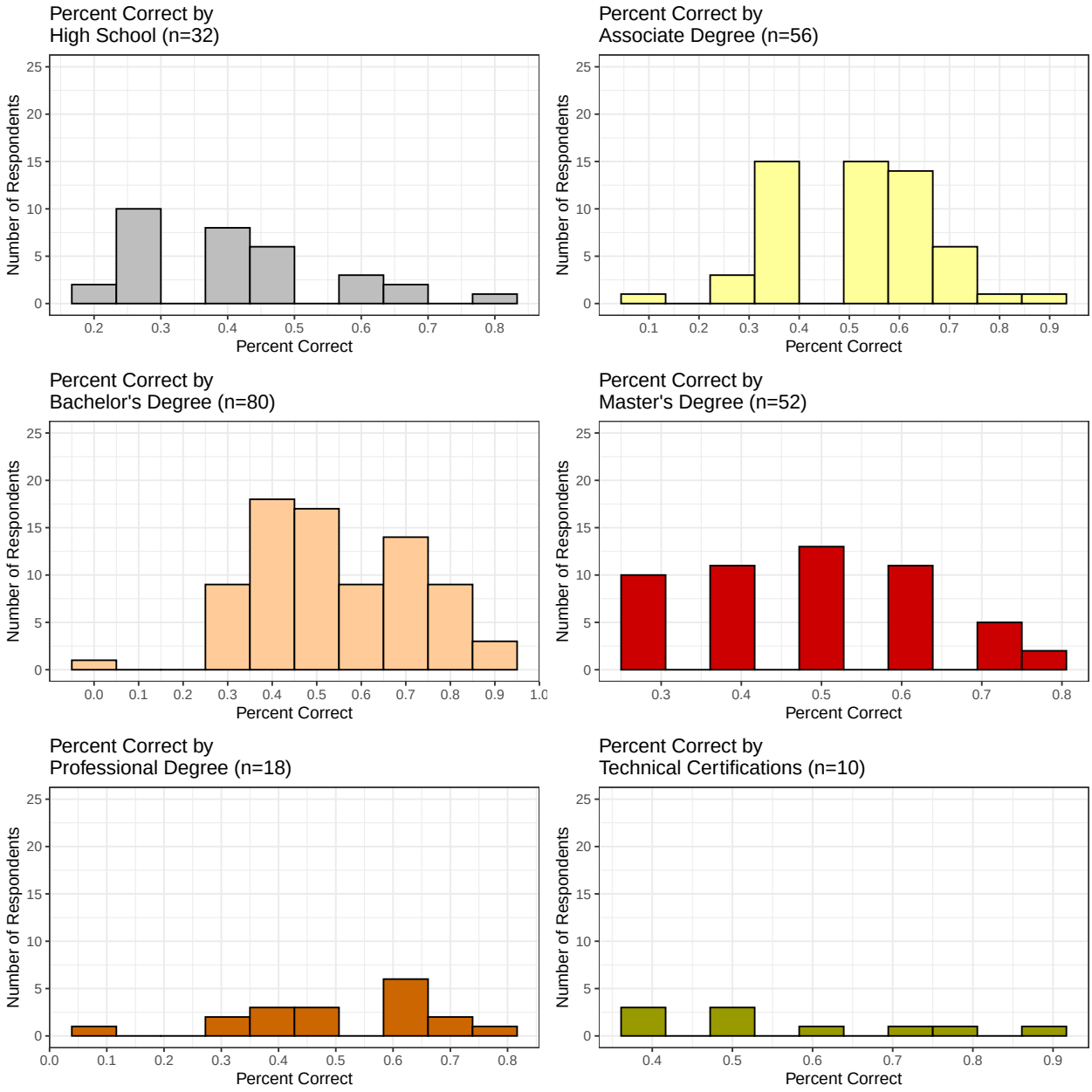

Percent Correct by Condition
Number of Respondents
Percent Correct
Condition
Passive x No Timer
Passive x Timer
Active x No Timer
Active x Timer
Passive x No Timer
Passive x Timer
Active x No Timer
Active x Timer

We used a multi-factorial design because those that are in the active condition would be forced to spend more time reading the page before they could progress through the training. We see that respondents in the passive learning condition without a timer spend far less time on each training page than those in the active learning condition. Yet when we include whether they were in the no timer or timer condition, we do not see any statistical significance. This indicates that spending more time reading or interacting with phishing training does not improve accuracy on identifying emails as legitimate or phishing.

We evaluated whether people spent the same amount of time in the passive vs. active training pages. If they were spending the same amount of time regardless, these results could still be confounded with time spent even though we imposed a timer. We find that they are spending significantly less time in the passive and no timing condition in both the Fake Fee Training (p-value of 0.03) and the Google Doc Training (p-value of 0.05); this means that they are still retaining and understanding the information even though they are only reading the information. Even though they were generally spending less time than the other conditions in the Account Upgrade Training and the Career Services Training, they weren't significant.


Time Spent on Fake Fee Posting Training
Time (seconds)
250
200
150
100
50
0
Passive x No Timer
Passive x Timer
Active x No Timer
Active x Timer
Condition
Time Spent on Account Upgrade Training
Time (seconds)
150
100
50
0
Passive x No Timer
Passive x Timer
Active x No Timer
Active x Timer
Condition
Time Spent on Google Doc Training
Time (seconds)
150
100
50
0
Passive x No Timer
Passive x Timer
Active x No Timer
Active x Timer
Condition
Time Spent on Career Services Training
Time Spent on Page (seconds)
600
400
200
0
Passive x No Timer
Passive x Timer
Active x No Timer
Active x Timer
Condition

When we evaluate whether there are covariates which are absorbing some of the effects, we see that there is significance in the degree that the subjects are working towards. However, we see that working towards a high school diploma score significantly worse in identifying the emails correctly when compared to those working towards an associate degree (p-value<0.01), a bachelor's degree (p-value<0.01), master's degree (p-value=0.02), and technical certifications (p-value=0.03).



We see 31 strange responses. These primarily come from participants who gave "gibberish" responses to our final question "Please let us know any comments you have." and those who classified all the emails as "legitimate" or "phishing." Due to small cell sizes, we were unable to perform a chi-square test with good estimates (19 responses in the passive learning condition and 12 responses in the active learning condition).

Conclusions

Due to the lack of significant findings, we cannot say that having an interactive training about phishing in cybersecurity would be beneficial. Instead, our results suggest that the overall level of education may have an impact on identifying emails, but this cannot be ascertained without further research. It is still important to educate people about phishing and cybersecurity in general, and it may be that we should just provide higher education to more people.

> "I think that this was very helpful and made me aware of scammers." - A survey respondent's comment about our survey

Limitations and Future Enhancements

There were several limitations in our study. We noticed that there are response discrepancies, a very limited training on phishing, and we may not have captured the impulsivity when readers initially see a tantalizing email, especially as they were not expecting it. We also do not know why there is differential attrition, but we suspect it may be due to going over quota. We did not find any records that went over quota, which is strange because there were some records that came in after the last completed record we received.

We see that there is a discrepancy between the age and gender responses from Pure Spectrum's report compared to the ones we asked in Qualtrics. This is common in survey responses, and it would be better if we could have an independent data source verify the true age, but we do not have access to such data.

Our phishing training was very limited. We only told the survey takers what to watch out for in phishing emails, and not what an actual recruiter solicitation or job offer looks like in the treatment. Thus, the survey takers could have been primed to think that all the emails we presented were phishing emails; however, this was not reflected in the data and very few marked all the emails as legitimate (5 responses) or all as phishing (3 responses); we were unable to perform a chi-square test with numbers this low. Additionally, there are unique and shortened links in recruiter solicitation and interview emails that are unique (e.g., calendly invitations). If the advice is to not click on links, then it is not useful. Perhaps results would have been different if we were to include more varied training on what recruiter solicitation emails look like, especially if they are not from well-known companies.

We may not have impulsivity when respondents see an email with a tantalizing offer. We tried to capture it by asking whether respondents are job hunting and including it as a covariate, but respondents know this is a research study and we provided training that hopefully made them stop and analyze emails carefully. A field study where a university IT department sends out a fake phishing email in multiple waves may better capture whether students fall for employment phishing scams or not. If they were told that it was a phishing attempt when they fall for it on one wave and they should be wary in the future, it would be interesting to see if they fall for it on a subsequent wave.