Effect of Learning Critique Styles on Learning Outcomes

Fall 2020 - W241 Final Report

Dahler Battle, Guy El Khoury, Jane Hung, Julian Tsang

Load Data

```
d_respondents <- fread('datatable_clean_survey_responses_v2.dta')
d_respondents[ , US_Dummy := ifelse(country == "United States", 1, 0)]
```

Abstract

Background

Research Question

Our research highlights the broad field of research around the role of feedback on performance. Successful feedback mechanisms and improved performance are likely highly correlated with each other. However it is too broad of a question for an experiment to point to a causal claim. Exogenous factors such as the learning environment, the learner's psychological mentality, or the type of task being taught may come into play in an non-experimental analysis.

Additionally, this question can be broken down to ask what types of feedback are helpful. Specific types of feedback may be better than others and some may actually be detrimental to one's performance (due to increased stress or mental strain. This could be attributed to several lines of reasoning. As such, a well designed experiment is necessary to find a true (if any) causal effect on learning outcomes.

The scope of our experiment is, as a result, intentionally narrow to measure the effect of different types of feedback on recognizing if an x-ray has healthy lungs or lungs with pneumonia. It is a binary outcome is simple and makes many elements consistent throughout for most participants (i.e. the computer-based learning environment, the feedback types, and the question being asked are the same throughout the program).

Hypothesis

The research question in this experiment attempts to answer the following question:

> *What type of feedback (positive reinforcement, negative reinforcement, self-reflective, etc.) leads to the largest improvements in individual performance on simple, recognition-based tasks, if any?*

We are testing the null hypothesis that feedback does not lead to better outcomes. That the ATE for all feedback groups and those that receive a placebo will equal 0.

A related follow-up question addresses:

> *Does more frequent feedback yield higher task performance?*

We anticipate that more feedback touchpoints may be associated with better individual performance because the receiver has more insight into how to improve and is able to calibrate to meet and surpass previous performance thresholds.

Experimental Design

Overview

This design is a different in differences experiment. Participants completed the three party survey in one sitting. The random assignment occurs after the first round of questions and the core analysis compares the difference in scores between the first iteration (pre-treatment) and the combined second and third iteration scores (post-treatment).

In this experiment, participants will view a set of X-Ray slides. Each slide contains an X-Ray image of a patient's lungs. The participant will have to determine if the patient's lungs are healthy or have pneumonia. Responses and timings will be recorded. Three rounds will create an answer set of 30 images (10 X-Ray images x 3 Rounds). Each of these intervention type, while limited in scope to the X-Ray recognition task, is meant to replicate a style of feedback (as seen in "What it Reflects" below). Participants will be randomly assigned the following control or treatments, with two, one minute breaks in between sessions.

- *Control* - Subject watches a pharmaceutical video and is asked how the video makes them feel. This replicates the experience of someone without any form of internal or external feedback.
- *Self Reflective Treatment* - Subject shown the last round's images, their answer, and the correct answer. They are then asked to reflect in two sentences how they can improve.This reflects someone who does not receive feedback from others but thinks critically about their own performance and how to improve.
- *Positive Reinforcement Treatment* - Subject shown the images of the last round's healthy lungs only and is asked to study those images for 1 minute. This reflects someone who is only told the positive aspects of their performance.
- *Negative Reinforcement Treatment* - Subject shown the images of the last round's pneumonia filled lungs only and is asked to study those images for 1 minute. This reflects someone who is only told the negative aspects of their performance.
- *Specific Feedback Treatment* - Subject shown the last round's images, their answer, and the correct answer. They are then given medical textbook info about how to spot pneumonia. This reflects a situation where someone who is given expert-driven advice on how to accomplish a task.

Project Timeline

The project was conducted on the following timeline:

Experiment Ideation & Design	*Trial Survey*	*Survey Period*	*Data Collection & Analysis*	*Final Presentation*	*Final Report*
Oct. 28 - Nov. 5	Nov. 6 - 8	Nov. 9 - 14	Nov. 15 - 30	Dec. 8	Dec. 15

Enrollment and Recruitment Process

Subjects were recruited through Mechanical Turk and were properly incentivized to complete the survey by receiving $1 upon successful completion. MechanicalTurk lists the survey in a pool of others and payouts were given by the research team after successful completion of the survey. We ended up receiving 447 survey submissions. Since we charged too high of a price point per survey, we were able to receive all of these responses in a matter of 72 hours. This may have worked in our favor by mitigating time-series related effects in the resulting data, however included several drawbacks mentioned later in the paper.

Subjects were mostly from the United States (225) and India (115). There were more males that participated in the study (207) than females (143).

Communication and Measurement Tooling

The experiment recruited participants from Mechanical Turk, who were then given a link to the survey on Qualtrics. They were asked to enter their MTurk Worker ID and start the survey. The entire experiment flow was then run through Qualtrics. Every final participant was recruited through MTurk and did not rely on personal connections.

The survey was compatible with both mobile and desktop applications. This helped reduce the barrier to entry for the survey. To prevent participants from rushing through answers, we put timings on the responses and essentially required each subject to complete the survey in one sitting.

Exclusion and non-interference

ADD HERE

Randomization

Participants recruited were randomly assigned to each of the 5 groups based on randomization logic pre-built on the Qualtrics system. Randomization occurred through the Qualtrics system after the first pre-treatment phase and split the remaining responses evenly between the four treatment groups and the control group. The Qualtrics Flow can be seen below.

NOTE: Need to replace local path name with github path name

In future iterations of the experiment, researchers may want to have a much larger control group to eliminate concerns over statistical power (as discussed in Part VII, Limitations and Future Enhancements).

Covariate Balance Checks

We performed visual covariate balance checks on the survey data as it relates to gender, age range education, and country. We additionally performed a Chi Squared test to test for independence within each of these categories. None of the Chi-Squared tests were significant at the p = .05 level, signaling that is no relationship between these covariates and the treatment and control assignment groups.

```
create_heatmap <- function(var1, var2) {
  ### Create a heatmap for a table of frequencies between two variables ###
  df <- data.frame(table(var1,var2))

  ggplot(df,aes(x=var1,y=var2)) +
    geom_tile(aes(fill=Freq,color=Freq),show.legend=FALSE,alpha=.8) +
    geom_text(aes(label=Freq)) +
    theme(axis.text.x = element_text(angle = 90)) +
    scale_fill_continuous(high = "darkslategray4", low = "powderblue")
}

# check balance between genders
gender_chiqq <- chisq.test(d_respondents[ , table(Assignment_Group, Gender)])

create_heatmap(var1 = d_respondents$Assignment_Group,var2 = d_respondents$Gender) +
  xlab('Assignment Group') +
  ylab('Gender') +
  labs(title = 'Contingency table between gender and assignment group',
       caption = paste0('Assuming gender distributions are the same among assignment groups, a chi-squa
                        round(gender_chiqq$parameter,4),' \ndegrees of freedom ', 'yields p=',
                        round(gender_chiqq$p.value,4),
                        ', suggesting that there is no relationship between gender and assignment group
  theme(plot.caption = element_text(hjust = 0))
```

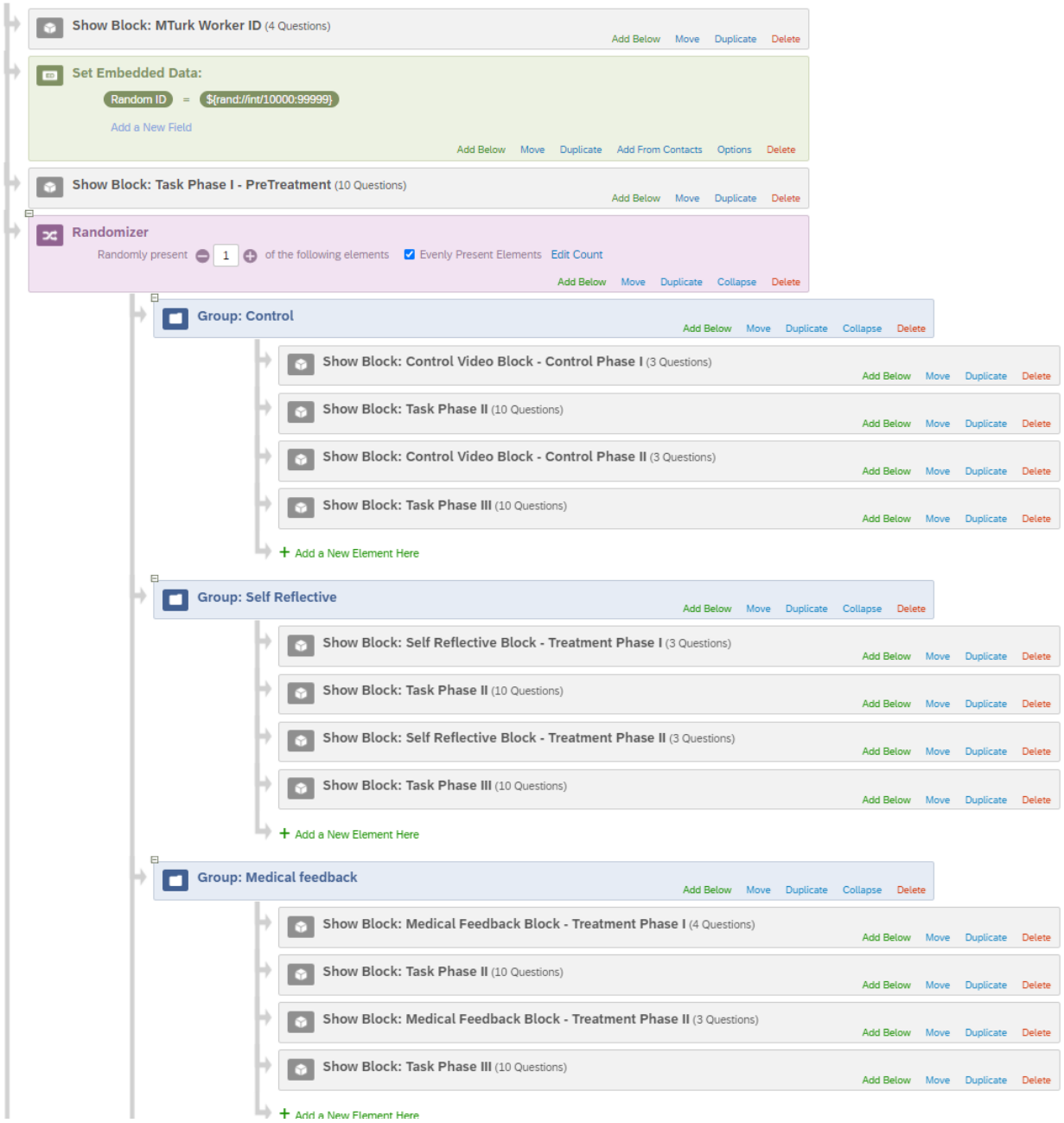
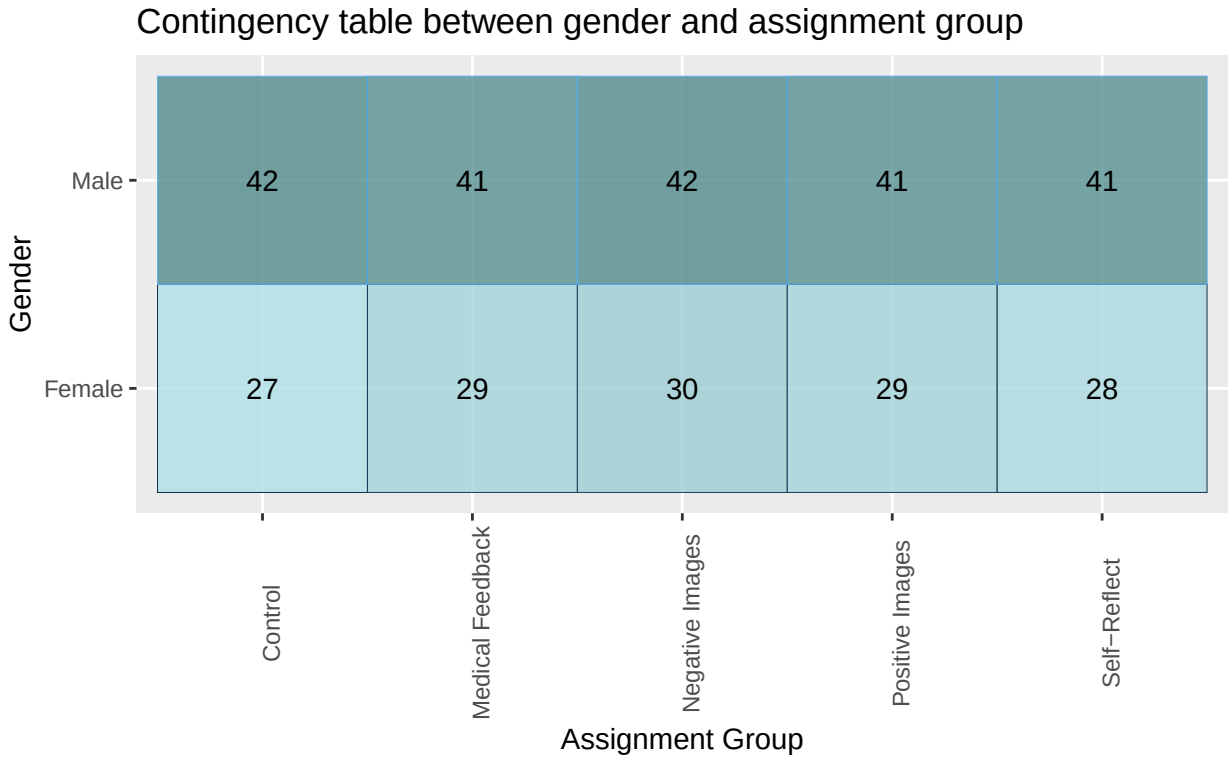
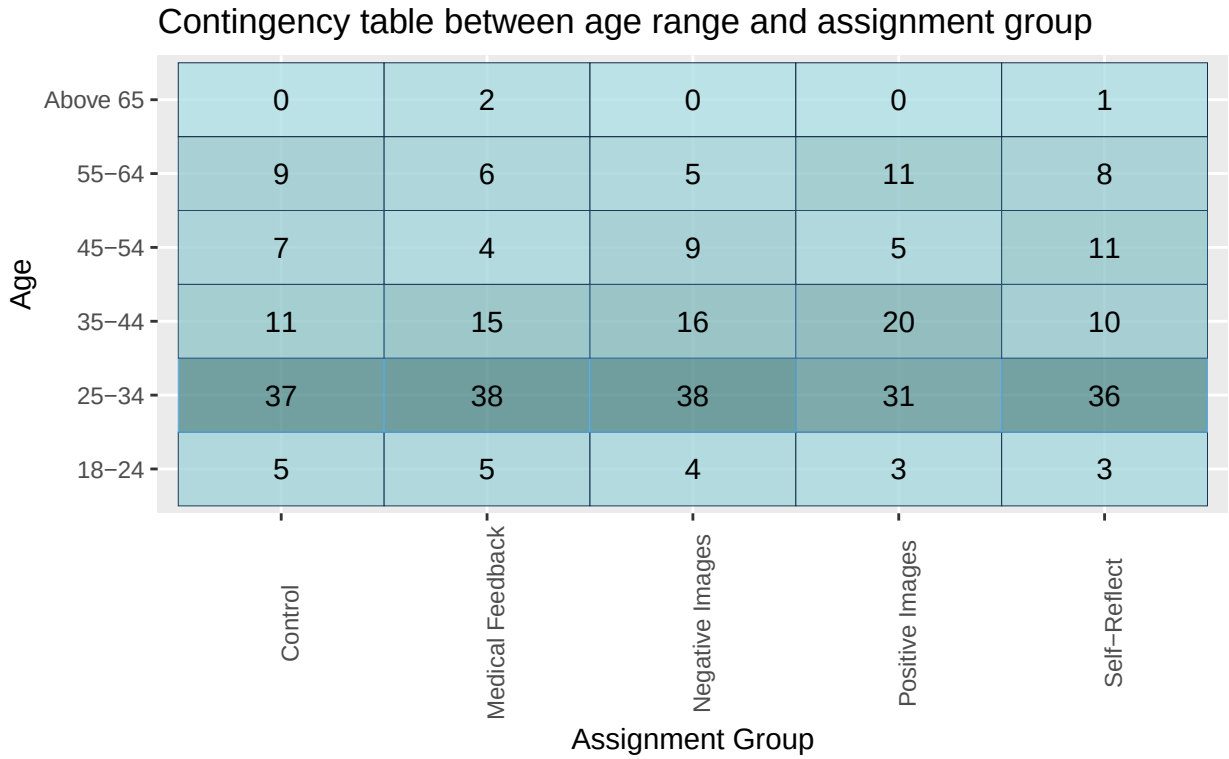


Figure 1: Qualtrics Flow



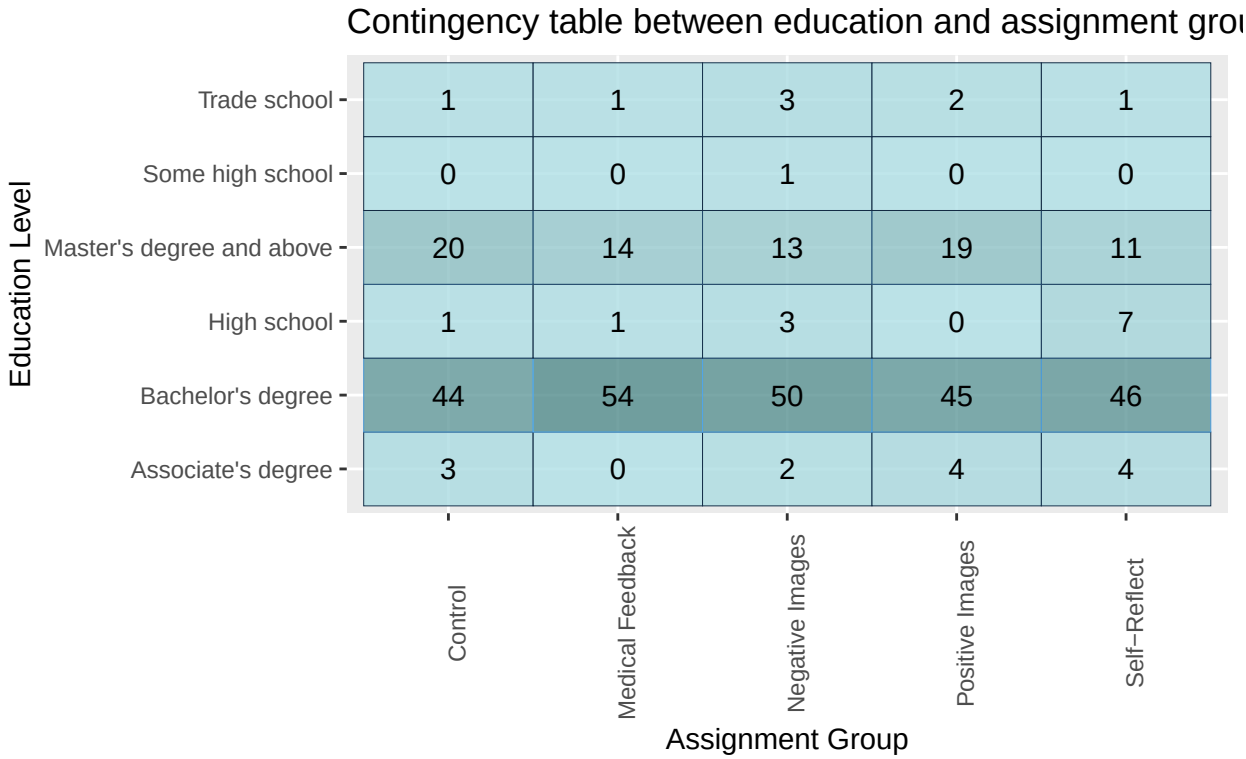
Assuming gender distributions are the same among assignment groups, a chi−squared test for independ
degrees of freedom yields p=0.9981, suggesting that there is no relationship between gender and assign
significance level of 0.05.

```
# check balance between age ranges
age_chisq <- chisq.test(d_respondents[ , table(Assignment_Group, Age_Range)],simulate.p.value = TRUE)

create_heatmap(var1 = d_respondents$Assignment_Group,var2 = d_respondents$Age_Range) +
  xlab('Assignment Group') +
  ylab('Age') +
  labs(title = 'Contingency table between age range and assignment group',
       caption = paste0('Assuming age distributions are the same among assignment groups, a chi-squared
                        round(age_chisq$p.value,4),
                        ', suggesting that there is no relationship between age and assignment groups a
  theme(plot.caption = element_text(hjust = 0))
```



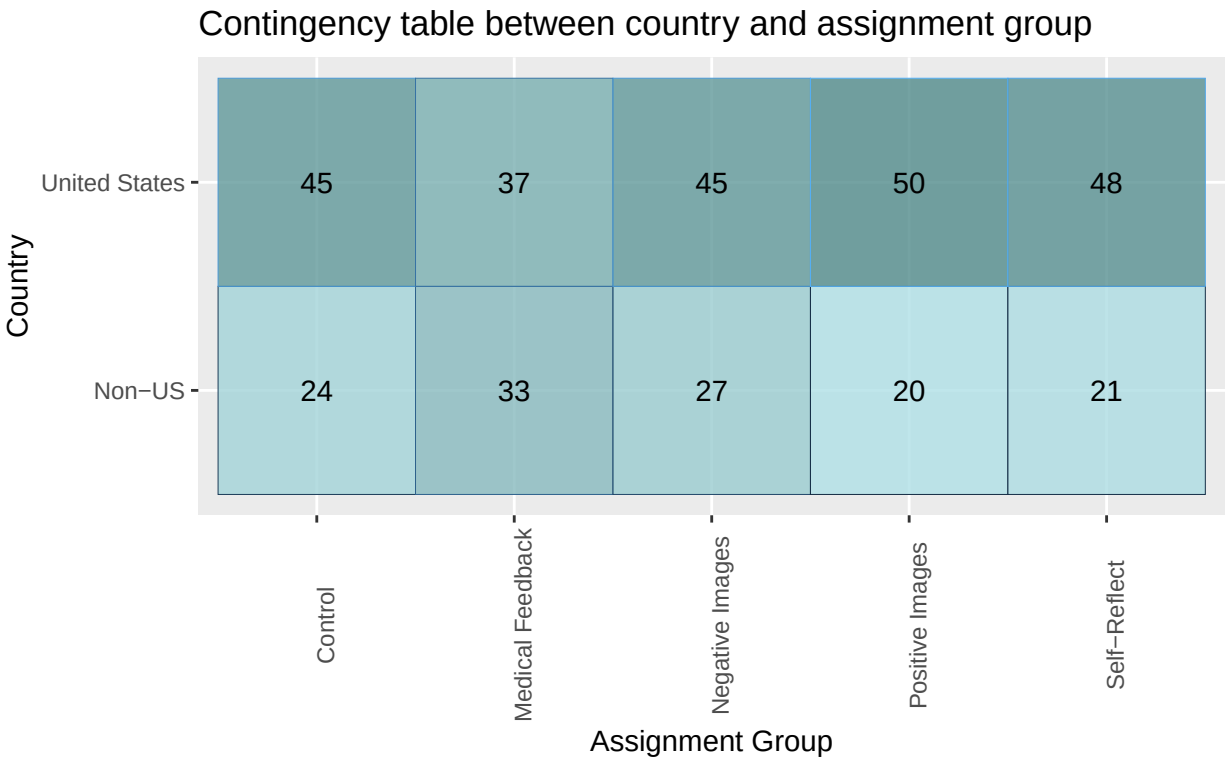
Assuming age distributions are the same among assignment groups, a chi−squared test for independe
Carlo simulation yields p=0.5027, suggesting that there is no relationship between age and assignmen
significance level of 0.05.

```
#check balance between education levels
edu_chisq <- chisq.test(d_respondents[ , table(Assignment_Group, Education_Level)],simulate.p.value = TR

create_heatmap(var1 = d_respondents$Assignment_Group,var2 = d_respondents$Education_Level) +
  xlab('Assignment Group') +
  ylab('Education Level') +
  labs(title = 'Contingency table between education and assignment group',
       caption = paste0('Assuming education distributions are the same among assignment groups, a chi-sq
                        round(edu_chisq$p.value,4),
                        ', suggesting that there is no relationship \nbetween education and assignment g
  theme(plot.caption = element_text(hjust = 0))
```



Assuming education distributions are the same among assignment groups, a chi−squ
independence with Monte Carlo simulation yields p=0.072, suggesting that there is n
between education and assignment groups at a significance level of 0.05.

```
# check balance between US and non-US respondents
us_chisq <- chisq.test(d_respondents[ , table(Assignment_Group, US_Dummy)])

create_heatmap(var1 = d_respondents$Assignment_Group,var2 = d_respondents$US_Dummy) +
  xlab('Assignment Group') +
  ylab('Country') +
  scale_y_discrete(breaks=c("0", "1"),
                   labels=c("Non-US", "United States")) +
  labs(title = 'Contingency table between country and assignment group',
       caption = paste0('Assuming country distributions are the same among assignment groups, a chi-squa
                        round(us_chisq$parameter,4),' degrees of freedom ', 'yields p=',
                        round(us_chisq$p.value,4),
                        ', suggesting that there is no relationship between country and assignment \ngro
  theme(plot.caption = element_text(hjust = 0))
```



Assuming country distributions are the same among assignment groups, a chi−squared test for ind
4 degrees of freedom yields p=0.1647, suggesting that there is no relationship between country an
groups at a significance level of 0.05.

Observation and Outcome Measurables

The data we collected was exported directly from Qualtrics into a CSV file. Data was then cleaned in R and exploratory data analysis was performed to check out data points. In all, we collected the following categorical data:

- Metadata - Entry data such as start and end dates, IP Addresses, Locations, Duration, Survey Status (Finished, Incomplete)
- Demographic Data - Age Range, Education Level
- Assignment Group - Control, Positive Images, Negative Images, Self-Reflection, and Medical Feedback
- Responses - Survey responses for Task Phase 1 (questions 1 - 10), Task Phase 2 (questions 11 - 20), and Task Phase 3 (questions 21 - 30)
- Scores - Scores for Task Phase 1, Task Phase 2, Task Phase 3 (out of 10); treatment scores combining Task Phases 2 and 3 (out of 10); cumulative scores (out of 30)

Data Completeness

The experiment started off with 381 surveys sourced through MTurk. Off this initial batch of participants, some submitted multiple responses in order to try to take advantage of our higher than average survey price point. We included only their initial surveys, throwing out 4 responses. 4 had not done the survey but had only submitted a code. Additionally, 37 surveys had blatantly intentionally incorrect answers in one or more sections. This includes survey participants who marked responses in all one answer (e.g. all healthy) or alternating answers throughout the survey (e.g. healthy, pneumonia, healthy, pneumonia, etc.).

Out of this participant pool, we threw out 97 results. These results were thrown out for the following reasons:

1. Multiple entrants (n = 4): The research team's $1.00 per survey price point was too high. As a result, some participants tried to send in multiple survey responses to collect multiple payments. In these

instances we only paid for (and used) the first survey.

2. Incomplete surveys (n = 56): Some people started surveys but never finished them. This includes those who never completed the last step of the survey, by closing out their answers. These responses were thrown out and attrition is dealt with in the survey results shown below.
3. Clear non-compliance (n = 37): Some participants did not give honest effort on the survey and answered all true, all false, or all alternating responses. These were also thrown out of the analysis.

Attrition could be counted in two buckets. First, the 56 incomplete surveys were incomplete before random assignment. An additional 43 dropped off after random assignment. Of these, 21 of these participants had made 99% progress but had failed to close the survey. However, we treated all 43 surveys left incomplete after assignment as part of attrition. A flow diagram below shows the drop offs at each level of the experiment below:

NOTE: Need to replace local path name with github path name

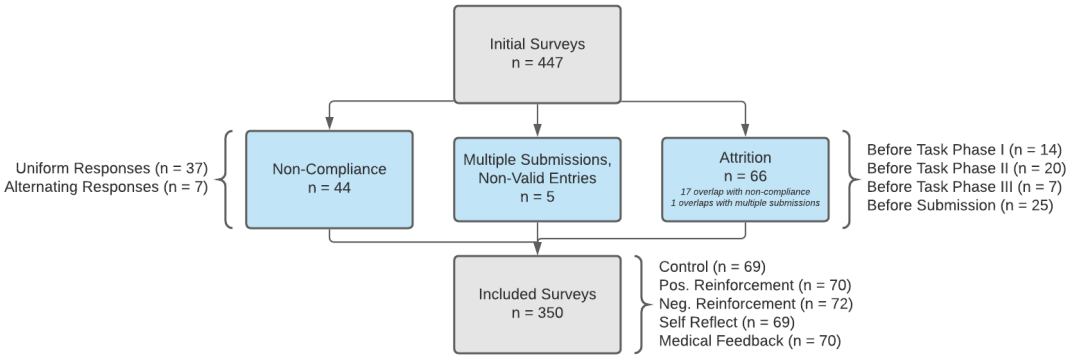


Figure 2: Survey Funnel Diagram

Results

Regressions

Power

```
power.t.test( delta = .05, sd = .16, sig.level = 0.05, power = 0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 161.711
##           delta = 0.05
##              sd = 0.16
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
power_n <- round(power.t.test( delta = .05, sd = .16, sig.level = 0.05, power = 0.8)$n)

d_respondents[, .N, by = .(Assignment_Group)]
```

```
##    Assignment_Group  N
```

```
## 1:  Negative Images 72
## 2:  Positive Images 70
## 3:     Self-Reflect 69
## 4:          Control 69
## 5: Medical Feedback 70
```

Power analysis shows that our groups did not have a large enough sample size required for each group. The number of subjects required is 162. Our group sizes for the control group, as well as the medical, positive, negative, and self-reflect treatment groups were 69, 70, 70, 72, and 69 respectively. This is primarily due to the fact that we charged too high of a price point per completed survey.

Conclusions

We can see from this study that we did find a statistically and practically significant impact of several styles of feedback as it relates to performance scores on the X-Ray lung health analysis. We can reject the null hypothesis that the scores in the feedback groups were the same as the scores in the placebo group. Breaking this down further, we see that (GO INTO THE INDIVIDUAL TREATMENT ATEs)

This experiment faces potential limitations when making more generalized conclusions about the effects of feedback on performance. For example, the experiment required analysis of a more simple, X-Ray analysis, which is not as complex of a task when compared to writing a paper or performing more quantitative analysis. Furthermore, the experiment's computer-facing setting may have impacted results. Subjects may not have spent as much time on the task as in a real life scenario. They certainly did not face the same time or social pressures of a real life task.

However, the study's conclusions gives us confidence that feedback positively affects performance in a meaningful way and more specifically targeted, informative feedback drives success. The effects of feedback on performance are signficant and merit additional study.

Limitations and Future Enhancements

The research design generated an output with limited power due to several factors. First, we handicapped the total amount of participants by offering too high of a price point for the survey. Our experiment offered a $1 price point per successful entry (limit of one entry per person). However we should have charged ~ $0.25. We also randomized our subjects equally between the four treatment groups and the one control group. In retrospect our group needed to allocate 50% of the participants to the control group and randomly assigned the rest in equal proportions to the four different treatment groups. These changes would have given the experiment an estimated 1400 participants, with XYZ in the control group and XYZ split evenly between the different treatment groups. Power for the experiment would have increased substantially and allowed for more meaningful outcomes.

Think about implementing a factorial design for our treatments so we can learn more from HTEs. is it the presence of images that improves scores? Is it the presence of a writing component? Is it the presence of reading?