Effect of Learning Feedback 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 <-
  d_respondents[!Amazon_Turk_ID %in% c(
    "A32K1M0A36EAK5",
    "A3I700VG8POSWH",
    "A3EPIT2P3ISA3K",
    "A390TP4DJI9P9W",
    "A3EPIT2P3ISA3K",
    "AUFLTHQAXWLH1",
    "AMZ1NNF182G8V",
    "A3BG20JPQLNKE1",
    "AVINXZZV3FNG7",
    "A3D4CITR1C9L3W",
    "A1BUYK6LXYWMLL",
    "A1FHRZXSE7XNJ4",
    "A12NQJV6TA5OWB",
    "AGCQCHHAQVA6R",
    "A119EX2L0DNN1B",
    "A3BPENSX5EVJ2H",
    "A371SNJNNUY9Z6",
    "A33YPBYXXZP3HY",
    "A3NYIJYBHAJ74V",
    "A2NGFU82LMJ80X",
    "A2J016DRTOBXWO",
    "A2GSZ3D2XXC533",
    "A2IGIOD74EPOEF",
    "A18WFPSLFV4FKY",
    "A1GMYDH5MKN105",
    "A8H1AW1NWPMAJ",
    "A39AAWF3F8OM6Q",
    "A26399B1QZ7XJJ",
    "A1NA92R0YGX32U",
    "A30UB0NUUWB3RM",
    "A16JX1MOPDCYDN",
    "A16X5FB3HAFCKN",
    "A211KGJ94WNFLN",
    "A1PGY59BR6C5BX",
    "AG5RF4UGQJ7A7",
```

```
    "A13T1DZG02HLNL",
    "A1WUFHQ1YGHK3C",
    "A1GNPQ5LM6YY5I",
    "A3EZ0H07TSDAPW",
    "ADLZLGHK0AEE6",
    "A1CF1W8CP0DHB0",
    "A3VP14XN3WUU0C",
    "A7VQQEIBSM9IU",
    "A8DER1QY96C5X",
    "A1M8MNKK8H5ZGW",
    "A1J3ICF1NZYFCR",
    "A34D5D6PU193AR"
  ), ]

nrow(d_respondents)
```

```
## [1] 333
```

Functions

Abstract

Background

Whether its the coach and player, teacher and pupil, or managers and direct reports, feedback likely plays an important role in delivering successful outcomes. All leaders are encouraged to give feedback while understudies are taught to receive critique openly. However, what is good feedback and how much of one's success on a given task be attributed to this feedback? Suprisingly few, well-developed experiments have been conducted to investigate this relationship. In this study, we seek to better understand if feedback truly influences successful outcomes and if different types of feedback lead to better outcomes than others.

Research Question

Our study highlights the broad field of research around the role of feedback on performance. Successful feedback is thought to lead to improved performance. 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, feedback comes in various forms, both positive and negative, internal and external. Some strategies may be better than others and others may actually negatively influence performance. As such, a well-designed experiment is necessary to find a true causal effect on learning outcomes (if any).

The scope of our experiment is, as a result, intentionally narrow to measure the effect of different types of feedback on task performance. In our design, we ask survey respondents to recognize if an x-ray image shows healthy lungs or lungs with pneumonia. This study introduces a novel concept to most, if not all subjects, requires strenous mental thought, and makes several extraneous elements consistent throughout the learning process (i.e. the computer-based learning environment, the feedback types, and the question being asked are the same throughout the program).

Hypothesis

Our study seeks to answer the following question:

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

We are testing the null hypothesis that the varying types of feedback do not lead to better outcomes. To generalize, we then test if the average treatment effect between those who receive any feedback and those who 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 will associate 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. However, it is unclear if the marginal gains from the second feedback loop will be as meaningful as the first.

Experimental Design

Overview

This design follows a difference-in-differences design and is implemented through regression adjustment. Participants completed a three-part survey in one sitting. The random assignment occurs after the first round of questions, which allows us to pre-screen for compliance. The core analysis compares the difference in scores between the first iteration (pre-treatment) and the second iteration (post first treatment) scores in order to test the immediate effects of feedback on performance. We further compare the first iteration scores with the third iteration (post second treatment) scores to understand the effect of repeated feedback.

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 (3 Rounds x 10 X-Ray images in each round). Participants will be randomly assigned to the following control or treatment groups, with two one-minute breaks in between sessions. Each intervention type, while limited in scope to the X-Ray recognition task, is meant to replicate a real-life style of feedback. The interventions are as follows:

- *Control* - Subject watches a pharmaceutical video and is asked how the video makes them feel. This replicates the experience of someone that does not receive any internal or external feedback.
- *Self Reflective Treatment* - Subject is shown the last round's images, their answers, and the correct answers. They are then asked to reflect in two sentences about 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 is 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 is 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 is shown the last round's images, their answers, and the correct answers. They are then given easy-to-digest information from a medical textbook on how to spot pneumonia. This reflects a situation where someone 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 (MTurk) and received $1 upon successful completion. Multiple entries from the same respondent were not permitted. Mechanical Turk 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 a relatively high 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 it also included several drawbacks mentioned later in the paper.**

Subjects were mostly from the United States (217) and India (106). There were more males that participated in the study (197) than females (136).

Communication and Measurement Tooling

The recruited Mechanical Turk participants were then given a link to the survey on Qualtrics. They were asked to enter their MTurk Worker ID and complete demographic questions before starting the survey. Friends and family were used to test the experiment flow, however none were known to have taken the full experiment, nor were part of our final analysis. The survey was compatible with both mobile and desktop applications. This helped reduce the barrier to entry for the survey. To help prevent non-compliance, we mandated timings on the treatment phases so that each subject fully received treatment.

Randomization

Since subjects were recruited from Mechanical Turk, we the experiment had access to a global pool of candidates. Then, participants 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. This randomization process is important so that treatment assignments are independent of subjects' potential outcomes. Furthermore, unaccounted-for covariates of the subject pool would not bias our estimate of the ATE.

The Qualtrics flow can be seen below.

Excludability and Non-Interference

This design also meets the excludability and non-interference assumptions needed to provide an unbiased estimate of the average treatment effect. Once a subject is assigned a treatment group, he or she receives a specific treatment for two separate times since treatment phases alternate with task phases 2 and 3. We meet the excludability assumption since outcomes are measured consistently through all task phases and for all assignment groups. Every task phase is scored on a scale from 1 to 10. Thus, what one subject scored in pre-treatment can be directly compared to what he or she scored in post-treatment. Furthermore, subjects are asked to essentially make diagnoses from looking at x-ray images. We believe that this is an esoteric topic, which would make it difficult for respondents to perform third-party research while completing the survey. However, we are better able to answer this subject by looking at the completion times below.

```
#n survey responses > 30 mins., take outlier out for analysis but discuss below
outlier <- round(d_respondents[`Duration (in seconds)` > 60*30, `Duration (in seconds)`/60/60],1)
completions <- d_respondents[`Duration (in seconds)` < 60*30]
```

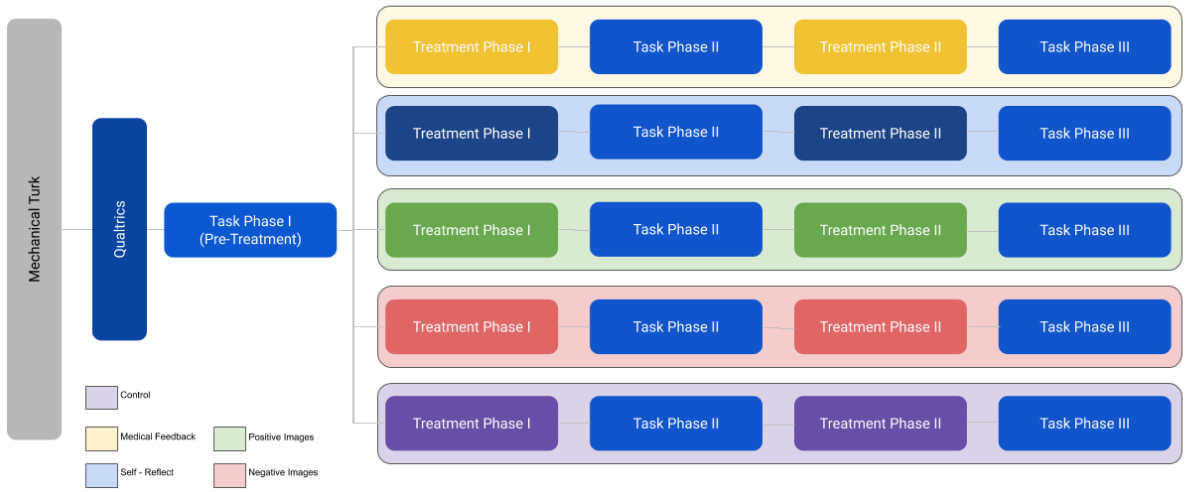
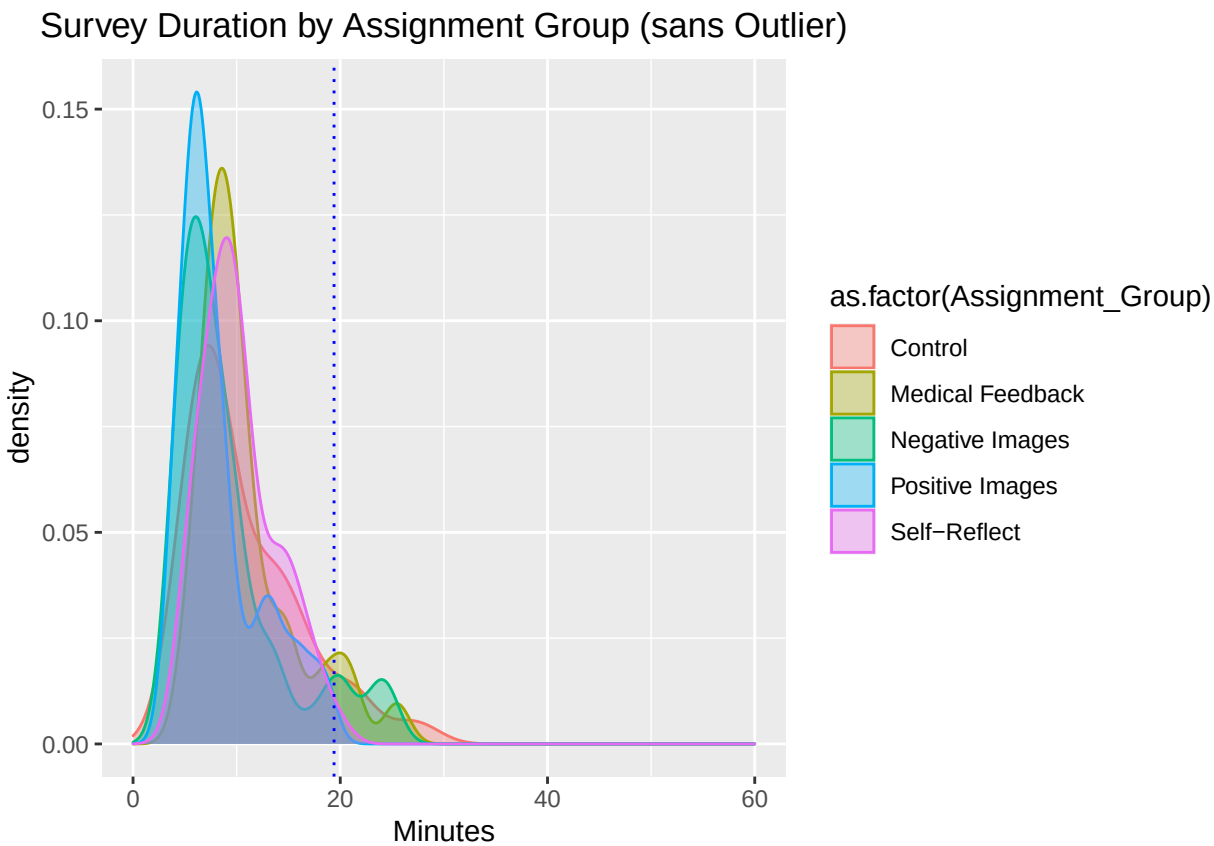


Figure 1: Qualtrics Flow

```
#95% of participants finished below this point in mins.
upper_cl <- completions[, round(mean((`Duration (in seconds)`)/60) + (2 *(sd(`Duration (in seconds)`)/6

#density plot of time completed by assignment group in mins.
ggplot(completions, aes(x=`Duration (in seconds)`/60, fill = as.factor(Assignment_Group), colour=as.fac
  geom_density(alpha = 0.35) +
  xlim(0,60) +
  ggtitle("Survey Duration by Assignment Group (sans Outlier)") +
  labs(x = "Minutes") +
  geom_vline(xintercept = upper_cl, linetype="dotted", color = "blue", size = 0.5) +
  theme(plot.title = element_text(hjust = 0.5))
```



We had one entry that took 4.9 hours to complete the survey. This could be due to research but is likely due to other factors such as just leaving the computer idle up for certain period of time. Eliminating this outlier, 95% of participants completed the survey in 19.4 minutes or less (6.4667 minutes or less per task phase). As such, subject driven, third-party research did not likely play a role in outcomes. The non-interference assumption is also met in this experiment since subjects are not aware of the treatments in other groups. They also do not know each other and cannot share about their treatment status with untreated subjects or vice versa.

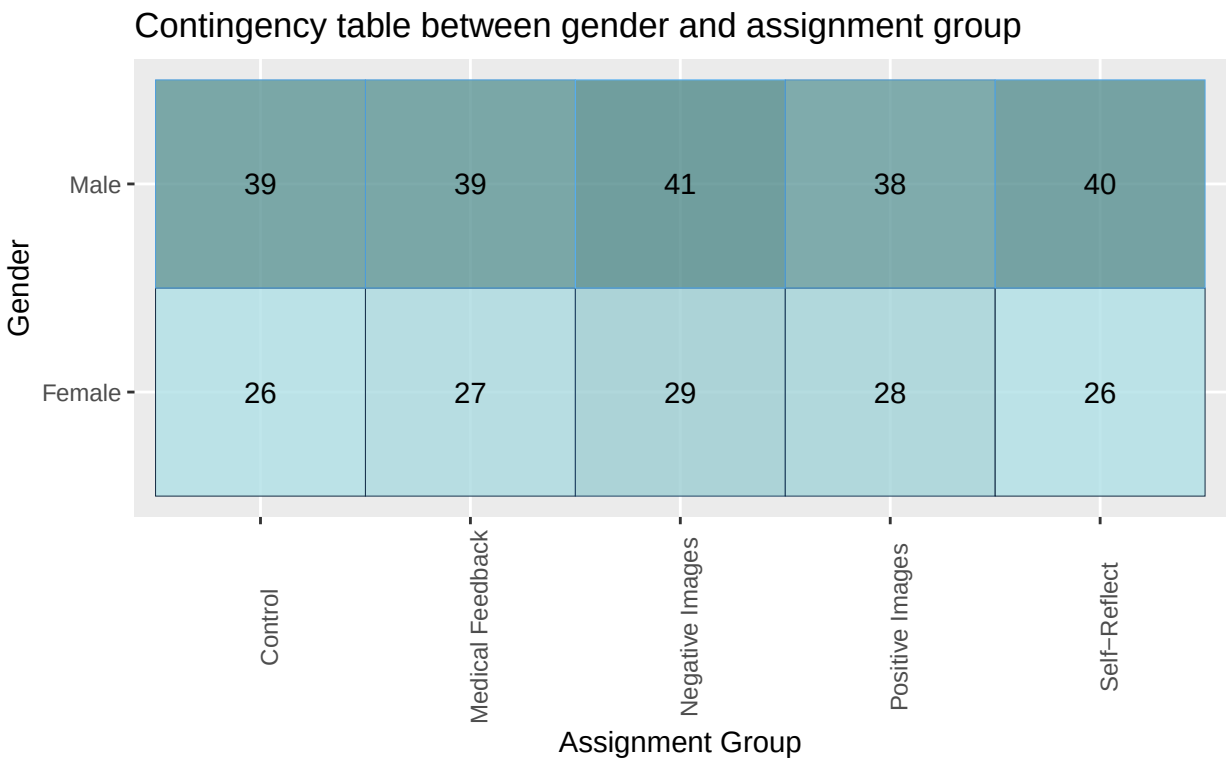
Covariate Balance Checks

We examined how well our randomization worked by checking that the proportion of individuals assigned to each group was similar. Furthermore, we performed visual covariate balance checks on the survey data as it relates to gender, age range, education, and country. We additionally performed Chi Squared Tests for Independence to test for independence within each of these categories. None of the Chi-Squared tests were significant at the p = .05 level, signaling that there is no relationship between these covariates and the treatment and control assignment groups. Proportions of each covariate were consistent across assignment groups.

```
# check balance between assignment groups
d_respondents[, .N, by = .(Assignment_Group)]
```

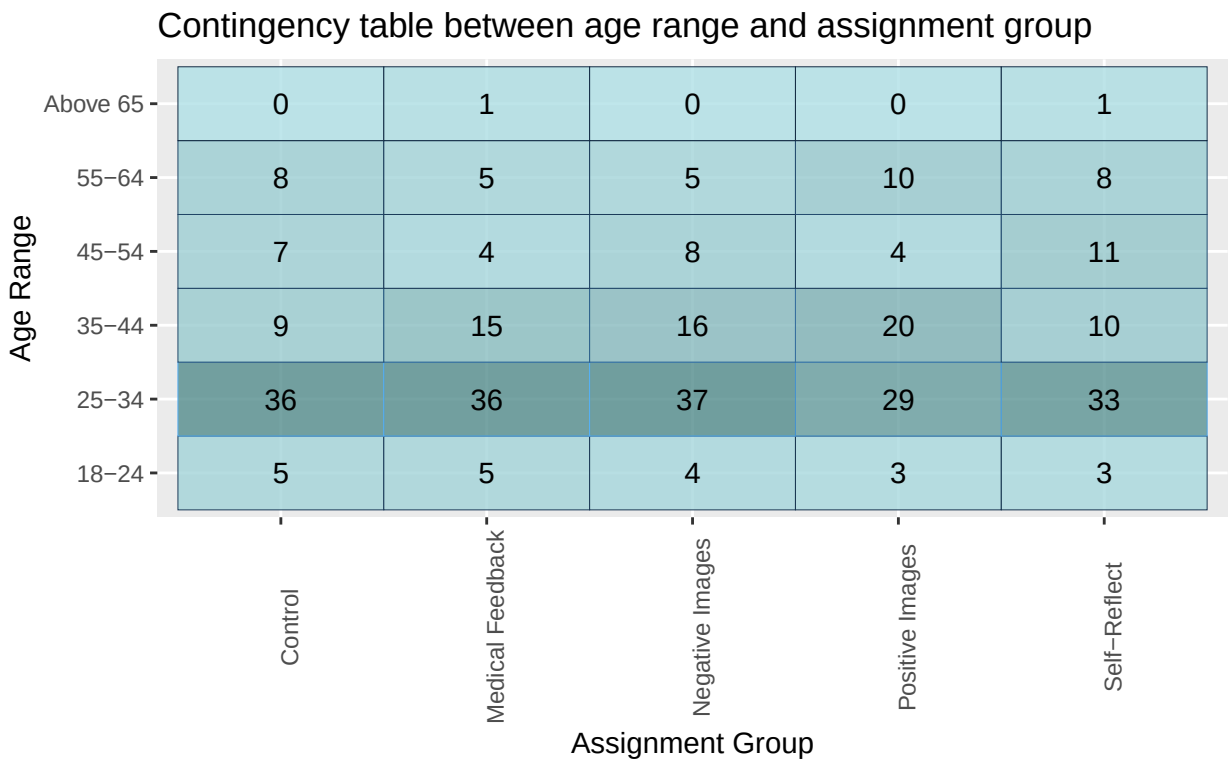
```
##    Assignment_Group  N
## 1:  Negative Images 70
## 2:  Positive Images 66
## 3:     Self-Reflect 66
## 4:          Control 65
## 5: Medical Feedback 66
```

```r
# check balance between genders
gender_chisq <- 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_chisq$parameter,4),' \ndegrees of freedom ', 'yields p=',
                        round(gender_chisq$p.value,4),
                        ', suggesting that there is no relationship between gender and assignment group
  theme(plot.caption = element_text(hjust = 0))
```



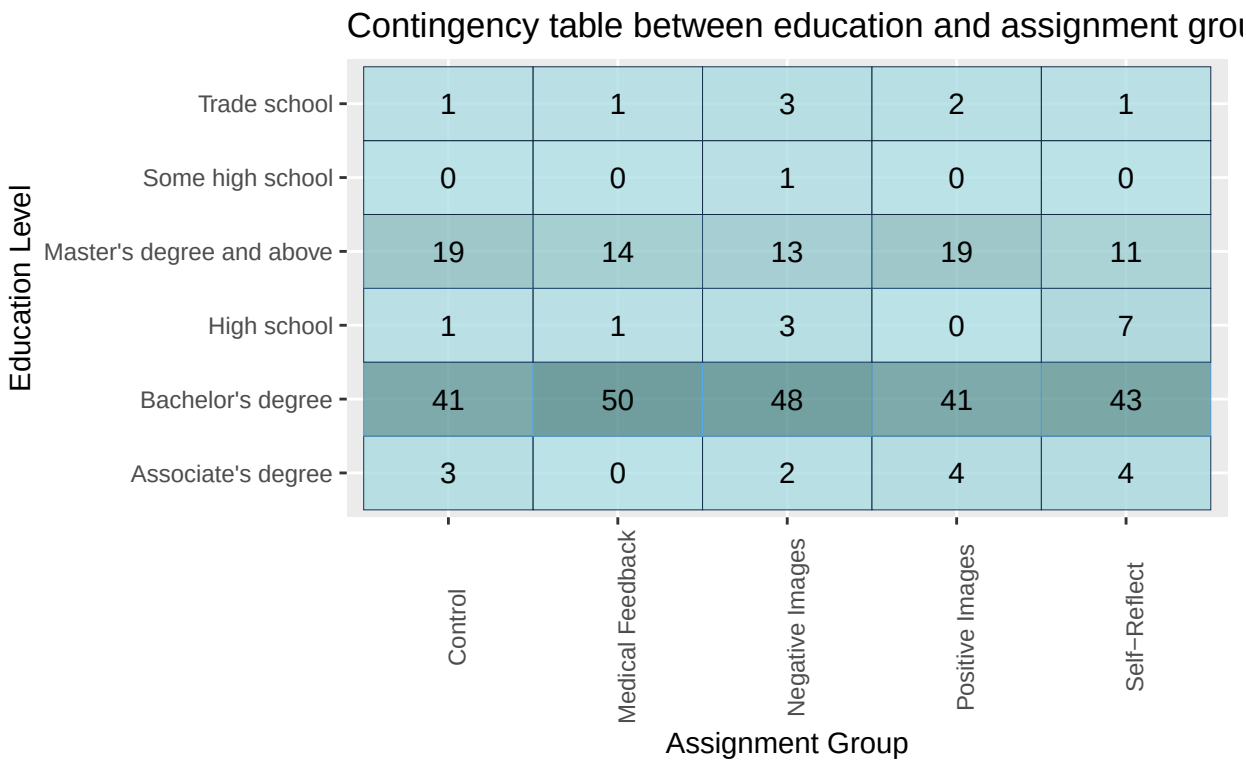
Assuming gender distributions are the same among assignment groups, a chi−squared test for independ
degrees of freedom yields p=0.9972, suggesting that there is no relationship between gender and assign
significance level of 0.05.

```r
# 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 Range') +
  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))
```

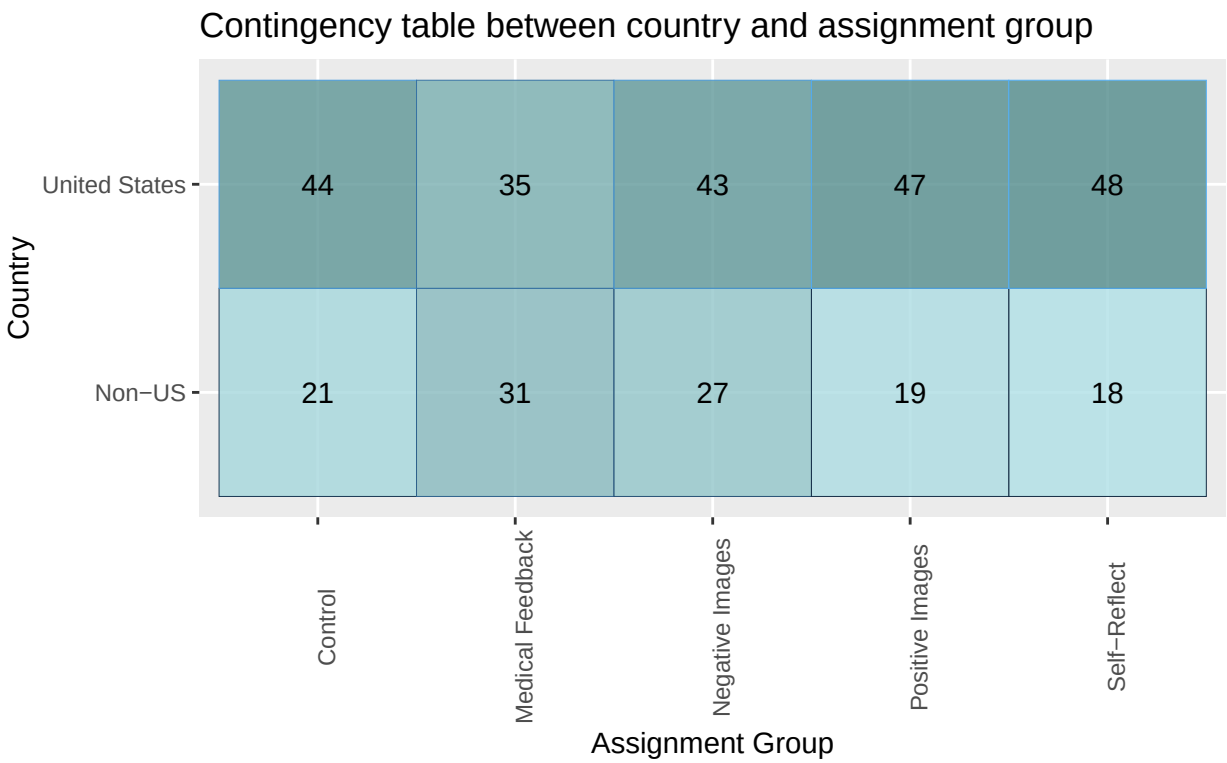


Assuming age distributions are the same among assignment groups, a chi−squared test for independe
Carlo simulation yields p=0.5447, suggesting that there is no relationship between age and assignmen
significance level of 0.05.

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

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-s
                        round(edu_chisq$p.value,4),
                        ', suggesting that there is no relationship \nbetween education and assignment g
  theme(plot.caption = element_text(hjust = 0))
```




```
# 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=0.1065, 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 better understand our 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, Gender
- 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)

Our outcome measurable follows a difference in differences deisgn. Scoring is based on number of questions a person gets right out of 10 questions per phase. This is then converted to a percentage value so that it will be easier to analyze regression results. So in this case, a 10 percentage point increase in performance would signify getting 1 additional question right. We will assess three main regressions with the following outcome variables: Task Phase 2 Scores and Task Phase 3 Scores.

We will focus on two major comparisons.

1. Control vs. All Treatment Groups: This compares people who receive the control with people who receive any form of feedback treatment.
2. Individual Treatment Effects: This second comparison focuses on comparing each individual treatment group with the control and with each other.

Data Completeness

The experiment started off with 381 surveys sourced through MTurk. Out of this participant pool, we threw out 97 results. These results were thrown out for the following reasons:

1. Clear non-compliance (n = 44): Some participants did not give honest effort on the survey and answered all "Normal", all "Pneumonia", or all alternating responses. These results were treated as instances of non-compliance and thrown out of the survey.
2. Multiple submissions and non-valid entries (n = 5): The research team's $1.00 per survey price point was relatively high. As a result, some participants tried to send in multiple survey responses to collect multiple payments or submit an invalid MTurk code (1 instance). In these instances we only paid for (and used) the first survey.
3. Incomplete surveys (n = 66): Some people started surveys but never finished. This includes those who never completed the last step of the survey by closing out their answers. These responses were thrown out and dealt with as instances of attrition.

Attrition occurred at several steps in the survey. 14 dropped off before Task Phase 1 while collecting demographic information and while entering the MTurk code (did not receive treatment assignment). 20 dropped the survey during the 10 image set in Task Phase 1 or during the first treatment phase. 7 dropped off during Task Phase 2 or during the second treatment phase. 4 dropped out during Task Phase 3 and 21 of these participants had made 99% progress but had failed to close the survey. However, we treated all 66 of the aforementioned incomplete survey responses as part of attrition and were not part of our final analysis. A funnel diagram below shows the participant drop offs of each type and at each level of the experiment:

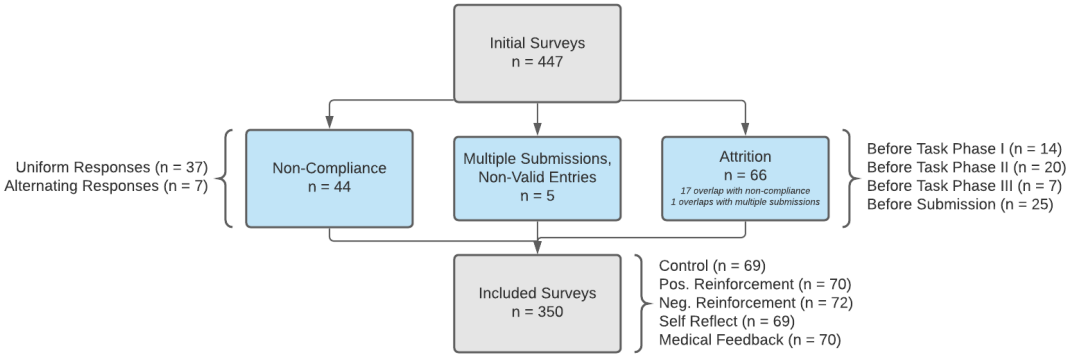


Figure 2: Survey Funnel Diagram Flow

Results

Overall, we have multiple ways we could have assessed this data based on our different treatment groups. We'll primarily focus on two major comparisons.

- Control vs. all treatment groups: This compares people who receive the control with people who receive any form of feedback treatment.
- Differences in individual treatment groups: The second comparison focuses on comparing each individual treatment group with the control and with each other.

Task Score Analysis

mpare task scores in different phases

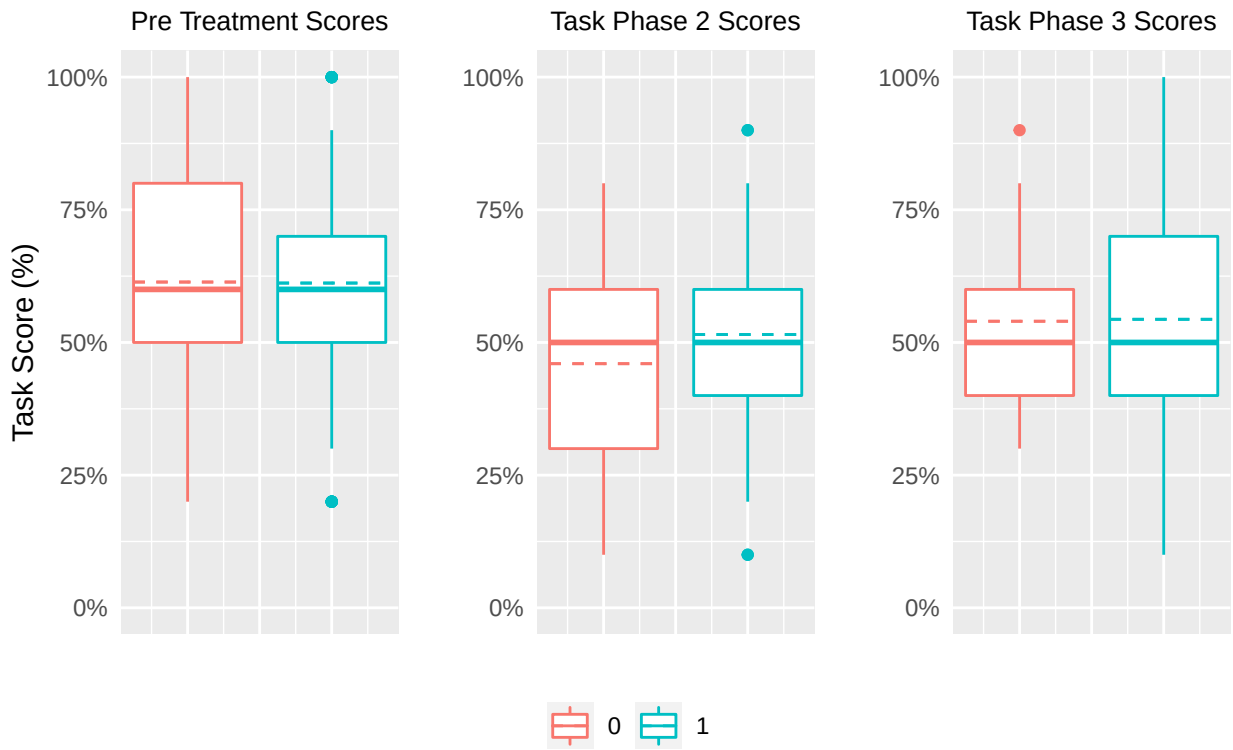
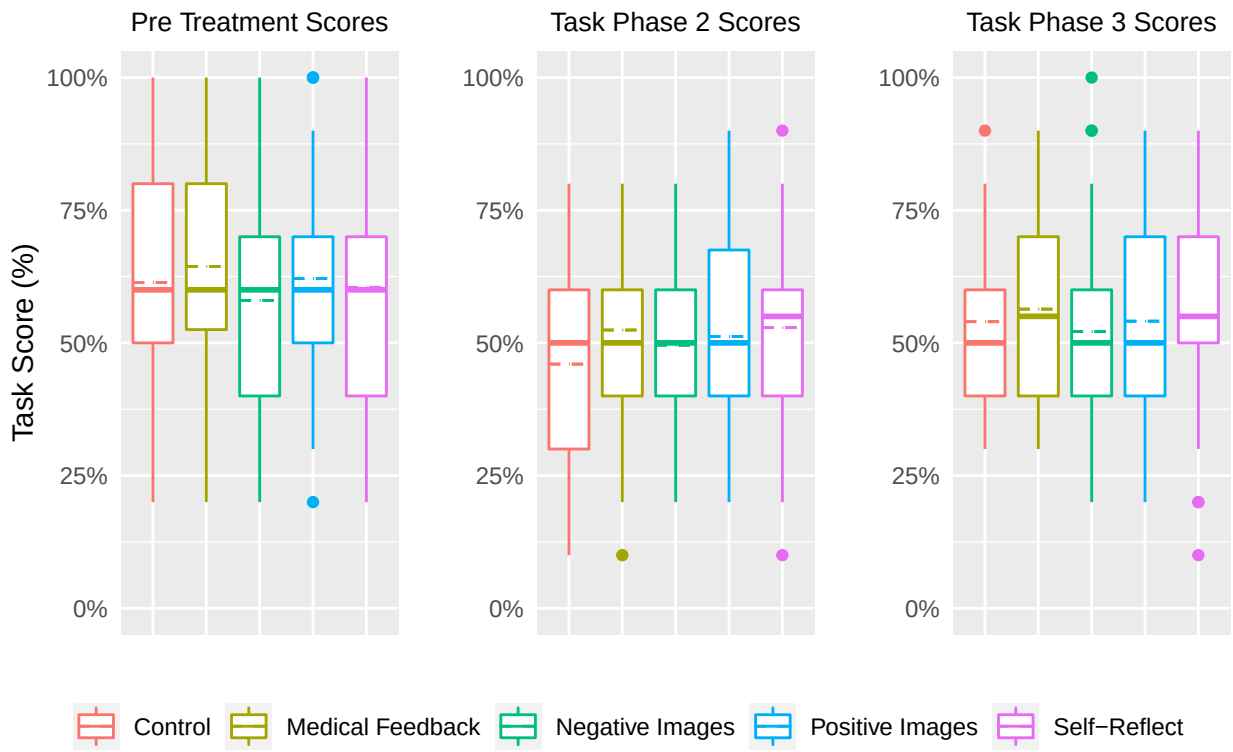
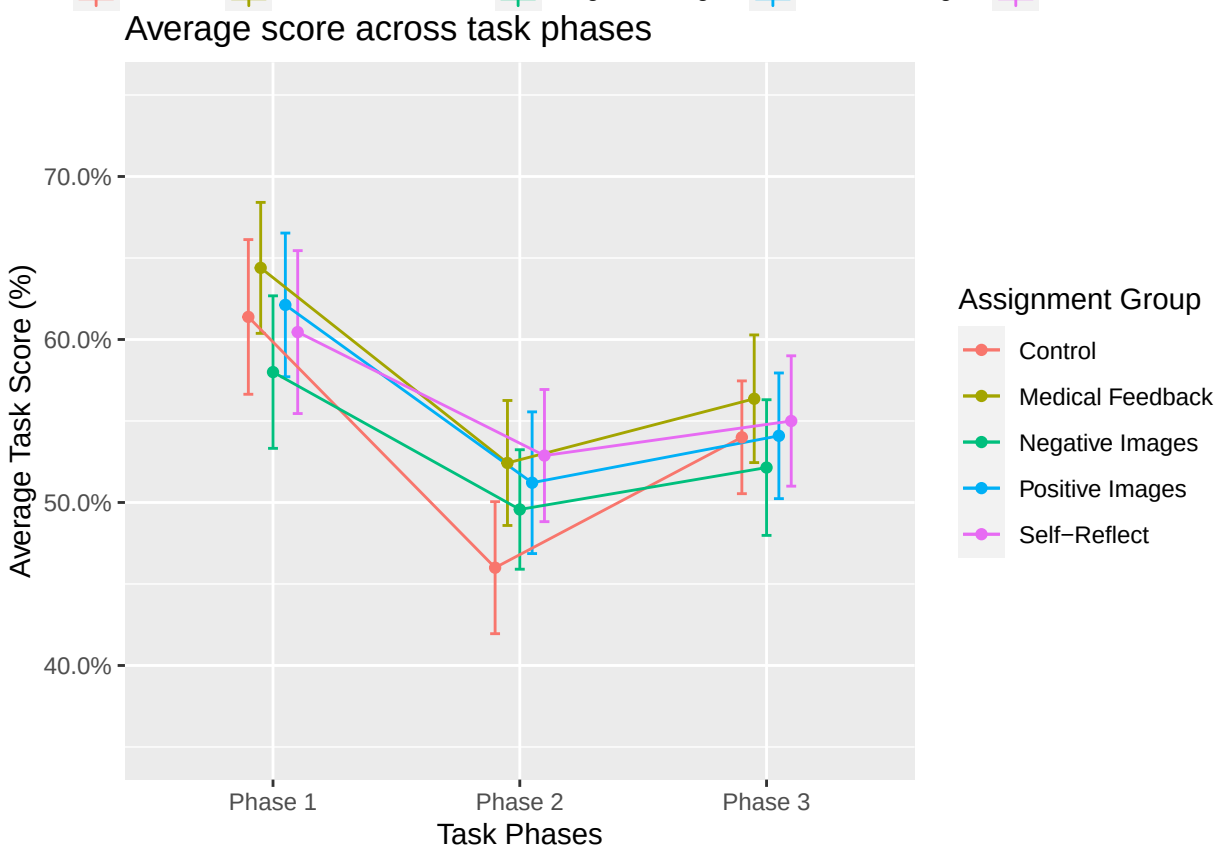


Table 2: Welch Two Sample t-test: `d_respondents[Treatment_Dummy == 0, TaskPhase1_Score]` and `d_respondents[Treatment_Dummy == 1, TaskPhase1_Score]`

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
0.07098	95.68	0.9436	two.sided	0.6138	0.6119

mpare task scores in different phases





Regressions

Table 3:

	Dependent variable:			
	Task Phase 2 Score			
	(1)	(2)	(3)	(4)
Any Treatment	0.055**	0.053**		
	(0.023)	(0.024)		
Medical Feedback			0.064**	0.059**
			(0.028)	(0.030)
Negative Images			0.036	0.044
			(0.028)	(0.028)
Positive Images			0.052*	0.049*
			(0.030)	(0.029)
Self-Reflection			0.069**	0.061**
			(0.029)	(0.031)
Task Phase 1 Score		0.241***		0.239***
		(0.048)		(0.049)
Male		−0.010		−0.010
		(0.018)		(0.019)
US		0.009		0.009
		(0.022)		(0.023)
Constant	0.460***	0.276***	0.460***	0.277***
	(0.021)	(0.074)	(0.021)	(0.074)
Education FE	No	Yes	No	Yes
Age FE	No	Yes	No	Yes
Observations	333	333	333	333
R^2	0.017	0.120	0.022	0.121
Adjusted R^2	0.014	0.081	0.010	0.074
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
318	8.147	NA	NA	NA	NA
315	8.133	3	0.01339	0.1728	0.9147

In the left table you see, we wanted to condense cells into assessing "any feedback" by creating a treatment dummy variable. We reasoned that in the real world, managers may have diverse ways of giving feedback, but at the end of the day, the direct reports are still receiving ways to understand their past performance and how to improve. Therefore, we want to roll up to a treatment dummy regression to test out this theory.

As we look at the second column of this table that contains our covariates (pre-treatment score, Gender, FE from education and age), we see that we experience a 5.1 percentage point increase in task performance when any feedback is given to the survey respondents, which is statistically significant given robust SE of 2.2. What is notable, though, is that adding in these covariates does not severely change our estimate for the effect of feedback and we do not see a marginal decrease in SE, so our estimate is no more precise when controlling for these other variables. As a gut check, we see that each 10% increase in Task Phase 1 scores is associated with a 2.4 percentage point increase in performance, which resonates with us; people who perform

well before feedback may also perform well after feedback.

Building off these results, we were further interested in exploring what type of feedback would yield the most positive impact on task performance. In this way, we were hoping to inform managers what type of feedback they should use with their direct reports. We believed that medical feedback would yield the most benefit because not only do you get information on what you got wrong but you also received expert opinion on how to properly assess the images. Abstracting this out to the real world, this would be akin to having a manager act as a mentor and using their experiences to enable your success. At a high level, we see that when people receive medical feedback, they experience a 5.5 percentage point increase in performance that is statistically significant, thereby confirming our hypothesis.

We hypothesized that the negative images feedback would fare the worst because we are only sharing their responses on the pneumonia images and whether they got them right or wrong. In this way, we wanted to simulate when a manager focuses on giving feedback only in abnormal situations. As a result, direct reports may have a poorer understanding of what "normal" or "good" looks like. Taking a look at that estimate, we see that people in the negative image feedback group have only a 3.9 percentage point increase that is not statistically significant, indicating that negative feedback was not helpful in improving performance.

Somewhat surprisingly, we found that people who were asked to self-reflect on their responses had a statistically significant 5.8 percentage point increase in performance. We were primarily interested in pursuing this type of feedback because it is a common personal growth technique to self-reflect that is touted in articles in HBR, Forbes, etc. In this way, we were able to confirm the positive effects of self-reflection; as a manager, you might encourage this behavior through incorporating self-assessments.

Lastly, was this necessary to blow out this analysis to the multiple treatment groups? An F-test suggests that expanding on the treatment groups as shown in the table on the right does not yield a model that better represents this data.

Table 6: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
318	7.982	NA	NA	NA	NA
315	7.953	3	0.02836	0.3744	0.7715

As a quick overview, we'd now like to assess Phase 3 results. As a reminder, this occurs after the subjects have received 2 rounds of treatment or placebo. We're anticipating that giving more feedback will yield even higher task performance scores compared to Phase 2, and we're hoping to understand if, as a manager, he/she should instantiate more touchbases to review performance.

What we see across the board though is that the effects of treatment are severely attenuated over time and with an additional round of feedback. For example, when assessing the effect of any feedback, there is a meager .2 percentage point increase in performance, which is not statistically significant.

This may be attributed to a number of things. For example, more frequent feedback during this short time span may be annoying to the receiver. The receiver may have then given much less attention to the feedback because they just received some critique fairly recently. On the other hand, a respondent paying close attention to this feedback may experience increased context switching, which may detract from completing the actual task.

As a conclusion, we see that feedback has immediate positive effects on performance, specifically critique that provides SME or is completed through a self-assessment. Although we did not see statistically significant effects from repeated feedback, this further may be attributed to how we conducted our study and the timespan allotted.

Table 5:

	Dependent variable:			
	Task Phase 3 Score			
	(1)	(2)	(3)	(4)
Any Treatment	0.004	0.004		
	(0.020)	(0.020)		
Medical Feedback			0.024	0.015
			(0.027)	(0.027)
Negative Images			−0.019	−0.013
			(0.028)	(0.027)
Positive Images			0.001	0.007
			(0.026)	(0.026)
Self-Reflection			0.010	0.006
			(0.027)	(0.028)
Task Phase 1 Score		0.154***		0.148***
		(0.048)		(0.048)
Male		−0.006		−0.006
		(0.018)		(0.018)
US		−0.007		−0.007
		(0.021)		(0.021)
Constant	0.540***	0.520***	0.540***	0.524***
	(0.018)	(0.065)	(0.018)	(0.065)
Education FE	No	Yes	No	Yes
Age FE	No	Yes	No	Yes
Observations	333	333	333	333
R^2	0.0001	0.082	0.007	0.085
Adjusted R^2	−0.003	0.042	−0.005	0.036
Note:				*p<0.1; **p<0.05; ***p<0.01

Exploratory Discussion - Noncompliance During Task Phases

Based on our definition of noncompliance, we do not consider noncompliance during the treatment phases. Although we have excluded noncompliers from our main regressions, for exploratory reasons, we can consider a scenario in which we do include them in our analysis. We start with examining the distribution of noncompliers across control and treatment groups. Interestingly, there are more noncompliers in the "Negative Images" treatment group than in any other group. We run a pair-wise proportions test to see whether the take-up rates across treatment groups are similar. Because we are making multiple comparisons, we use the Boneferroni correction and we seethere are no statistically significant differences between take-up rates across groups at the p = 0.05 level.

```
kable(d_respondents_noncompliers[ , table(Assignment_Group)])
```

Assignment_Group	Freq
Control	12
Medical Feedback	10
Negative Images	7
Positive Images	8
Self-Reflect	9

```
pairwise.prop.test(x = c(total_MedFeedback_compliers, total_selfreflect_compliers, total_positiveimages_
                total_negativeimages_compliers, total_control_compliers),
          n = c(total_MedFeedback_rows, total_selfreflect_rows, total_positiveimages_rows,
                total_negativeimages_rows, total_control_rows), p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using Pairwise comparison of proportions
##
## data:  c(total_MedFeedback_compliers, total_selfreflect_compliers, total_positiveimages_compliers, t
##
##   1 2 3 4
## 2 1 - - -
## 3 1 1 - -
## 4 1 1 1 -
## 5 1 1 1 1
##
## P value adjustment method: bonferroni
```

```
prop.test(x = c(total_MedFeedback_compliers, total_selfreflect_compliers, total_positiveimages_complier
                total_negativeimages_compliers, total_control_compliers),
          n = c(total_MedFeedback_rows, total_selfreflect_rows, total_positiveimages_rows,
                total_negativeimages_rows, total_control_rows))
```

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  c(total_MedFeedback_compliers, total_selfreflect_compliers, total_positiveimages_compliers, t
## X-squared = 1.7, df = 4, p-value = 0.8
## alternative hypothesis: two.sided
## sample estimates:
## prop 1 prop 2 prop 3 prop 4 prop 5
## 0.8684 0.8800 0.8919 0.9091 0.8442
```

Now that we have calculated the take-up rate, we can also calculate the Intent-to-Treat effect and the

Complier Average Causal Effect, which would undilute our treatment estimates for the various treatment groups in our main regression table. While the CACEs remain very similar to the original coefficients found in column 3 in Figure 3, we see that the CACE for the Negative Treatment group actually becomes negative.

```
treatment_groups <- c("Any Treatment", "Medical Feedback", "Negative Images", "Positive Images", "Self-
list_CACEs <- round(c(CACE_Any_Treatment, CACE_Medical_Feedback, CACE_Negative_Images, CACE_Positive_Ima

CACE_list <- cbind(treatment_groups, list_CACEs)
kable(CACE_list)
```

treatment_groups	list_CACEs
Any Treatment	0.0613
Medical Feedback	0.0766
Negative Images	0.0386
Positive Images	0.0647
Self-Reflect	0.0668

Exploratory Discussion - Noncompliance During Treatment Phases

The only ways we can plausibly detect signs of noncompliance during treatment phases is examining the amount of time spent each user spent on his/her respective treatment pages in the survey. However, determining a hard-set rule that funnels some respondents into noncompliers would also require major assumptions. Thus, our discussion here is only exploratory and does not help us to adjust our treatment estimates. This section is for informative purposes and may help us formulate a better design in the future. In the following figures, we plot the distribution of the times spent during Treatment Phase 1 and Treatment Phase 2 for the 350 respondents that were included in our main analysis.

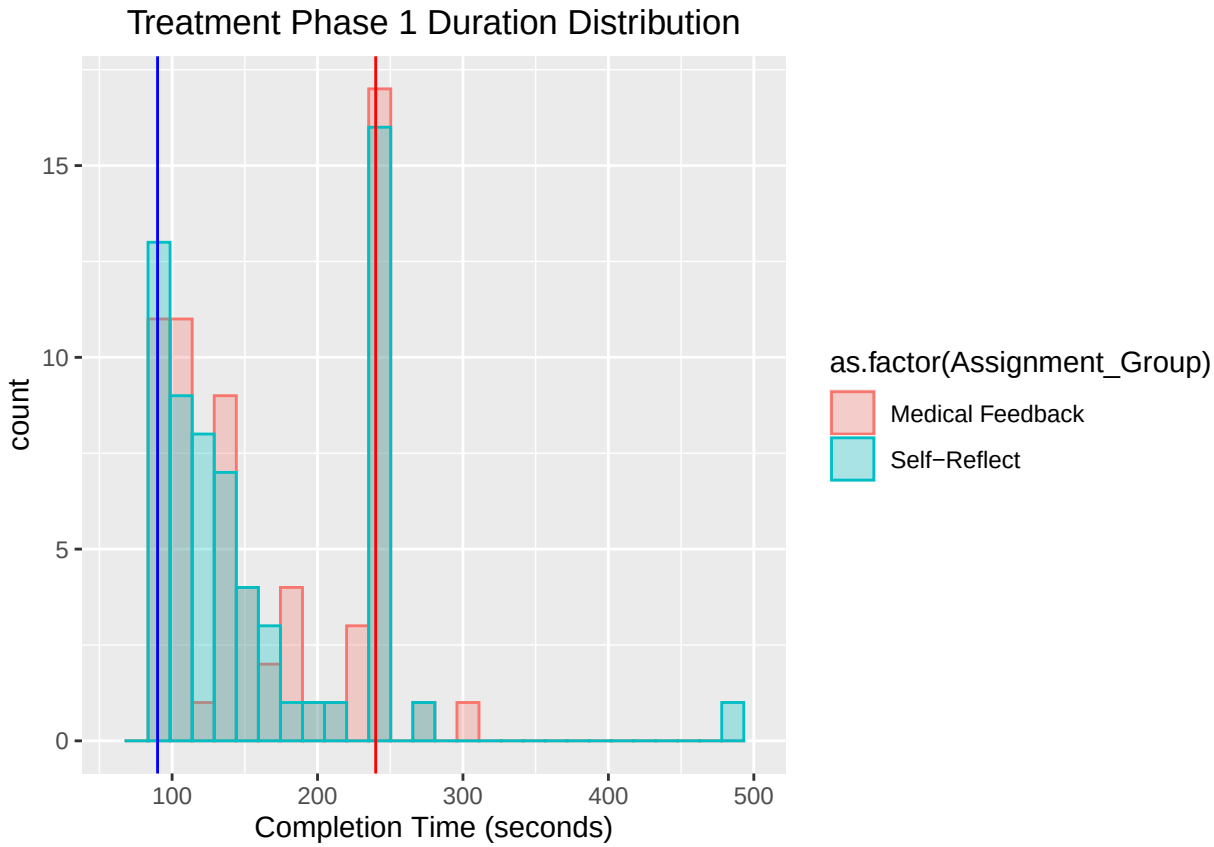
Both Medical Feedback and Self-Reflect treatment groups shared the same time constraints. We configured the settings so that respondents could proceed to the next page (leave the treatment phase) after 90 seconds have elapsed. The page would automatically advance to the next page after 240 seconds have elapsed. Thus, it is interesting to note that there were respondents who remained on the page beyond the 240-second time limit. We can visually see that respondents in the Medical Feedback group spent more time in Treatment Phase 1 than respondents in the Self-Reflect group did. There was a spike of respondents who exited treatment as soon as they were eligible to at the 90-second marker. Then, the number of people leaving treatment gradually decreased between the minimum and maximum time markers (reminiscent of the shape of a power-law distribution), followed by a second surge of leavers at the maximum time marker. Excluding those who stayed beyond the time limit, when conducting a T-Test comparing the difference in mean times between the two treatment groups, we find that the difference is not statistically significant at the $p = 0.05$ level.

During Treatment Phase 2, we notice that respondents in both treatment groups spent less time during treatment relative to Treatment Phase 1, as there was an increase of people in the Self-Reflect group who exited treatment as soon as they were eligible to at the 90-second minimum marker. Still, those in the Medical Feedback group stayed in the treatment phase longer than those in the Self-Reflect group did. Conducting a T-Test comparing the difference in mean times between the two groups during this treatment phase, we find that the difference is not statistically significant at the $p = 0.05$ level. However, the mean times in this treatment phase are less than the mean times in the previous phase, which brings up the question of whether there were more noncompliers who breezed through the second round of treatment. For both treatment groups, the differences between their respective mean times in Task Phases 1 and 2 are statistically significant (people clearly spent less time in the second round of treatment).

```
########## plot density distributions of timing for Treatment Phase 1: Medical Feedback vs Self-Reflect

ggplot(d_noncompliance_1[Assignment_Group %in% c("Medical Feedback", "Self-Reflect")], aes(x=Treatment_
```

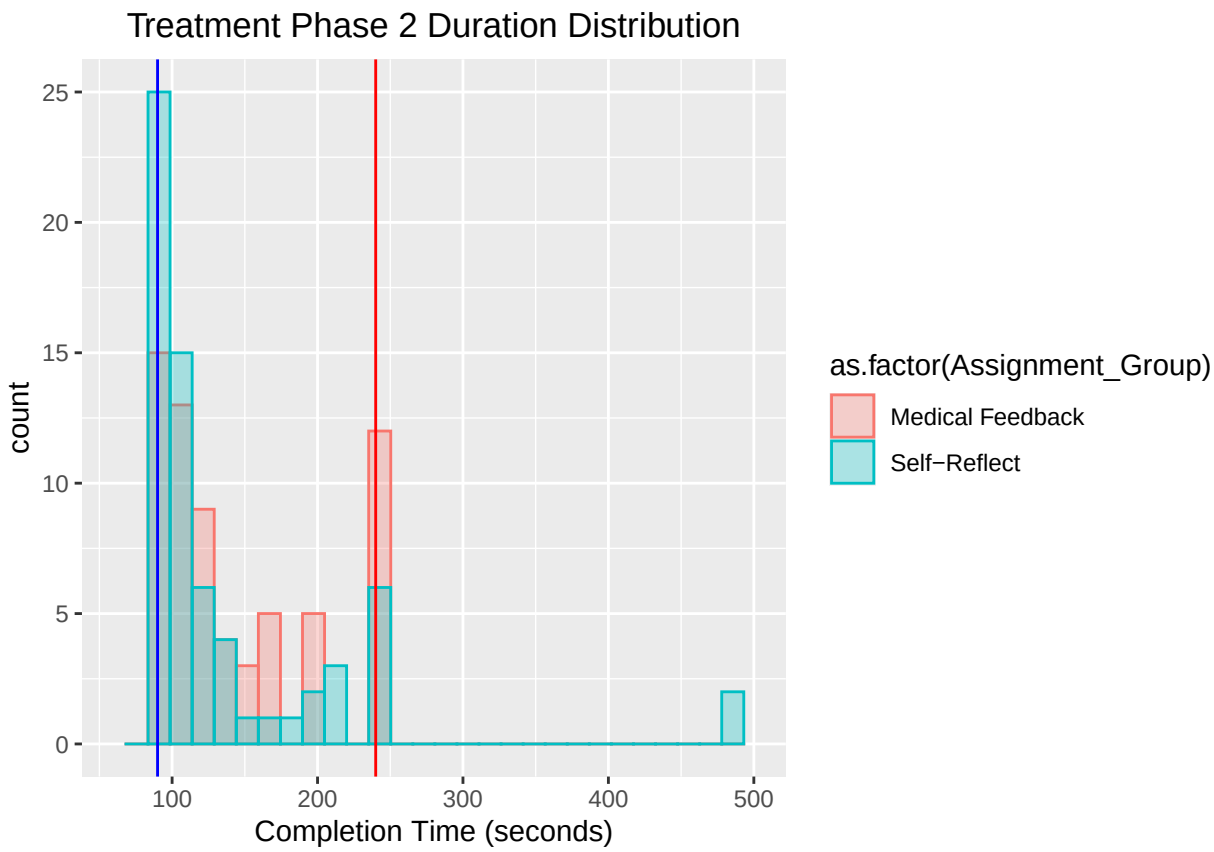
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
## Warning: Removed 4 rows containing missing values (geom_bar).
```




```
ggplot(d_noncompliance_2[Assignment_Group %in% c("Medical Feedback", "Self-Reflect")], aes(x=Treatment_
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 4 rows containing missing values (geom_bar).
```




```
t.test(d_noncompliance_1[Assignment_Group %in% c("Medical Feedback"), Treatment_Phase1_SubmitTime],
       d_noncompliance_1[Assignment_Group %in% c("Self-Reflect") & Treatment_Phase1_SubmitTime < 250, T
```

```
##
##  Welch Two Sample t-test
##
## data:  d_noncompliance_1[Assignment_Group %in% c("Medical Feedback"), Treatment_Phase1_SubmitTime] a
## t = 1.3, df = 127, p-value = 0.2
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.213 34.395
## sample estimates:
## mean of x mean of y
##     166.4     152.9
```

```
t.test(d_noncompliance_2[Assignment_Group %in% c("Medical Feedback"), Treatment_Phase2_SubmitTime],
       d_noncompliance_2[Assignment_Group %in% c("Self-Reflect") & Treatment_Phase2_SubmitTime < 250, T
```

```
##
##  Welch Two Sample t-test
##
## data:  d_noncompliance_2[Assignment_Group %in% c("Medical Feedback"), Treatment_Phase2_SubmitTime] a
## t = 1.9, df = 127, p-value = 0.06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4064 35.4432
## sample estimates:
## mean of x mean of y
```

```
##     144.7     127.2
```

```r
#statistically significant difference in time spent in treatment 1 and 2 for medical feedback
t.test(d_noncompliance_1[Assignment_Group %in% c("Medical Feedback"), Treatment_Phase1_SubmitTime],
       d_noncompliance_2[Assignment_Group %in% c("Medical Feedback"), Treatment_Phase2_SubmitTime])
```

```
##
##  Welch Two Sample t-test
##
## data:  d_noncompliance_1[Assignment_Group %in% c("Medical Feedback"), Treatment_Phase1_SubmitTime] an
## t = 2.1, df = 128, p-value = 0.03
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.593 41.816
## sample estimates:
## mean of x mean of y
##     166.4     144.7
```

```r
#statistically significant difference in time spent in treatment 1 and 2 for self-reflect
t.test(d_noncompliance_1[Assignment_Group %in% c("Self-Reflect"), Treatment_Phase1_SubmitTime],
       d_noncompliance_2[Assignment_Group %in% c("Self-Reflect"), Treatment_Phase2_SubmitTime])
```

```
##
##  Welch Two Sample t-test
##
## data:  d_noncompliance_1[Assignment_Group %in% c("Self-Reflect"), Treatment_Phase1_SubmitTime] and d_
## t = 2, df = 127, p-value = 0.05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.211 57.278
## sample estimates:
## mean of x mean of y
##     166.4     137.9
```

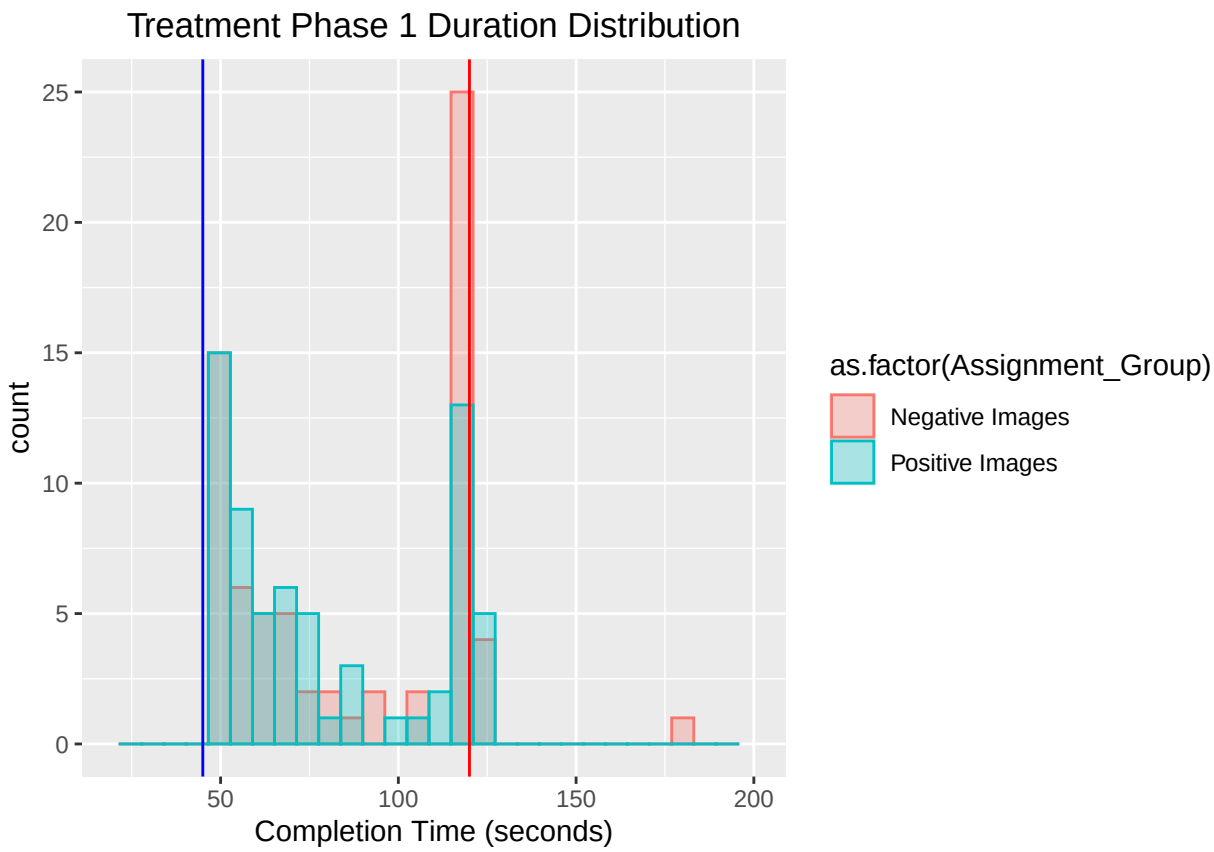
Both Positive Images and Negative Images treatment groups shared the same time constraints. We configured the settings so that respondents could proceed to the next page (leave the treatment phase) after 45 seconds have elapsed. The page would automatically advance to the next page after 120 seconds have elapsed. We noticed that there was one respondent who remained on the page beyond the 120-second time limit. Furthermore, we can visually see that respondents in the Negative Images group spent more time in Treatment Phase 1 than respondents in the Positive Images group did. We can confirm this with a T-Test, but the difference is not statistically significant.

In Treatment Phase 2, we see a similar increase in respondents who leave treatment at the minimum time marker. However, this increase is primarily driven by respondents in the Positive Images group as those in the Negative Images treatment group tended to stay in treatment longer. The difference in mean times between Positive Images and Negative Images groups are also statistically significant. In addition, respondents in the Positive Images group spent less time in the second round of treatment than the first, whereas there was no statistically significant difference in the amount of time that respondents in the Negative Images group spent in both rounds of treatment.

```r
########## plot density distributions of timing for Treatment Phase 1: Positive vs Negative Images

ggplot(d_noncompliance_1[Assignment_Group %in% c("Positive Images", "Negative Images")], aes(x=Treatmen
```

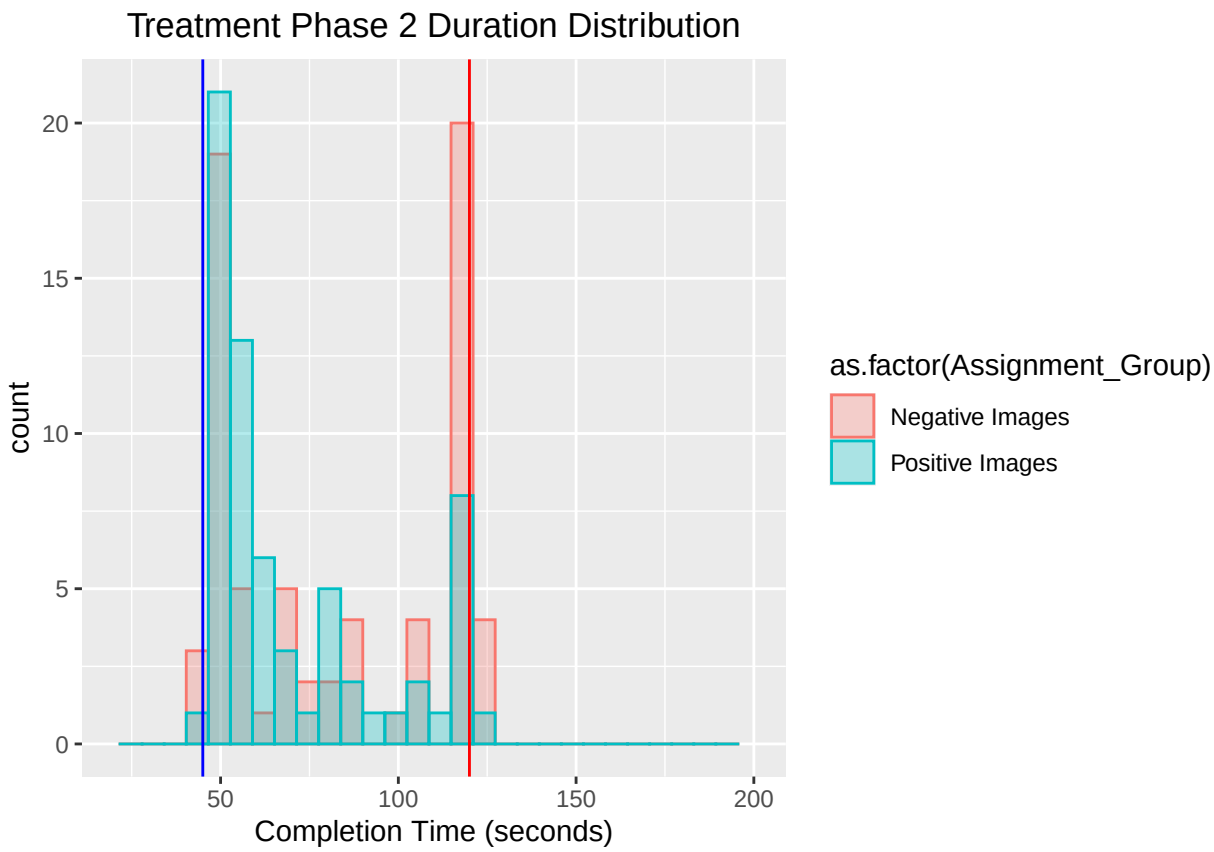
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```




```
ggplot(d_noncompliance_2[Assignment_Group %in% c("Positive Images", "Negative Images")], aes(x=Treatmen
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 4 rows containing missing values (geom_bar).
```




```r
t.test(d_noncompliance_1[Assignment_Group %in% c("Positive Images"), Treatment_Phase1_SubmitTime],
       d_noncompliance_1[Assignment_Group %in% c("Negative Images") & Treatment_Phase1_SubmitTime < 150
```

```
##
##  Welch Two Sample t-test
##
## data:  d_noncompliance_1[Assignment_Group %in% c("Positive Images"), Treatment_Phase1_SubmitTime] and
## t = -1.4, df = 133, p-value = 0.2
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -17.313   3.109
## sample estimates:
## mean of x mean of y
##     79.76     86.86
```

```r
t.test(d_noncompliance_2[Assignment_Group %in% c("Positive Images"), Treatment_Phase2_SubmitTime],
       d_noncompliance_2[Assignment_Group %in% c("Negative Images") & Treatment_Phase2_SubmitTime < 150
```

```
##
##  Welch Two Sample t-test
##
## data:  d_noncompliance_2[Assignment_Group %in% c("Positive Images"), Treatment_Phase2_SubmitTime] and
## t = -2.8, df = 131, p-value = 0.007
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -23.198  -3.823
## sample estimates:
## mean of x mean of y
```

```
##     69.70     83.21
```

```r
#statistically significant difference in time spent in treatment 1 and 2 for positive images
t.test(d_noncompliance_1[Assignment_Group %in% c("Positive Images"), Treatment_Phase1_SubmitTime],
       d_noncompliance_2[Assignment_Group %in% c("Positive Images"), Treatment_Phase2_SubmitTime])
```

```
##
##  Welch Two Sample t-test
##
## data:  d_noncompliance_1[Assignment_Group %in% c("Positive Images"), Treatment_Phase1_SubmitTime] and
## t = 2.1, df = 128, p-value = 0.04
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.6619 19.4473
## sample estimates:
## mean of x mean of y
##     79.76     69.70
```

```r
#no statistically significant difference in time spent in treatment 1 and 2 for negative images
t.test(d_noncompliance_1[Assignment_Group %in% c("Negative Images"), Treatment_Phase1_SubmitTime],
       d_noncompliance_2[Assignment_Group %in% c("Negative Images"), Treatment_Phase2_SubmitTime])
```

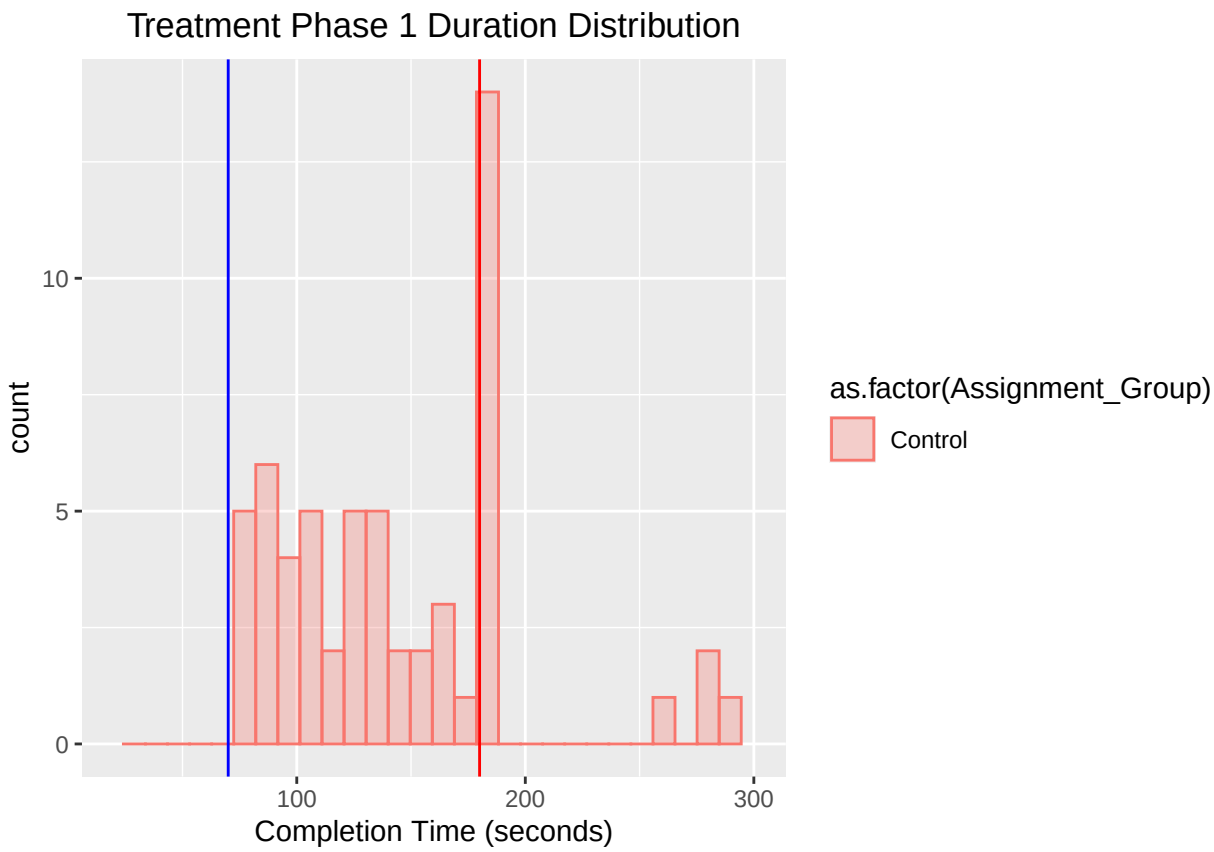
```
##
##  Welch Two Sample t-test
##
## data:  d_noncompliance_1[Assignment_Group %in% c("Negative Images"), Treatment_Phase1_SubmitTime] and
## t = 0.92, df = 138, p-value = 0.4
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.739 15.764
## sample estimates:
## mean of x mean of y
##     88.22     83.21
```

For the Control group, we configured the settings so that respondents could proceed to the next page (leave the treatment phase) after 70 seconds have elapsed. The page would automatically advance to the next page after 180 seconds have elapsed. We noticed that there were some respondents who remained on the page beyond the 180-second time limit. Furthermore, respondents in the Control group tended to spend more time in Treatment Phase 1, compared to those in any of the four treatment groups. Recall that in the treatment groups, we would observe a spike of respondents who left treatment at the minimum time marker in blue, which would gradually decrease as completion times increased. This would be followed by another spike at the maximum time marker. However, in the control group, we notice a different behavior in that the distribution of respondents between the minimum and maximum time markers would be relatively uniform.

During Treatment Phase 2, we also see that respondents spent less time in the second round of treatment relative to the first. This difference is statistically significant at the $p = 0.05$ level.

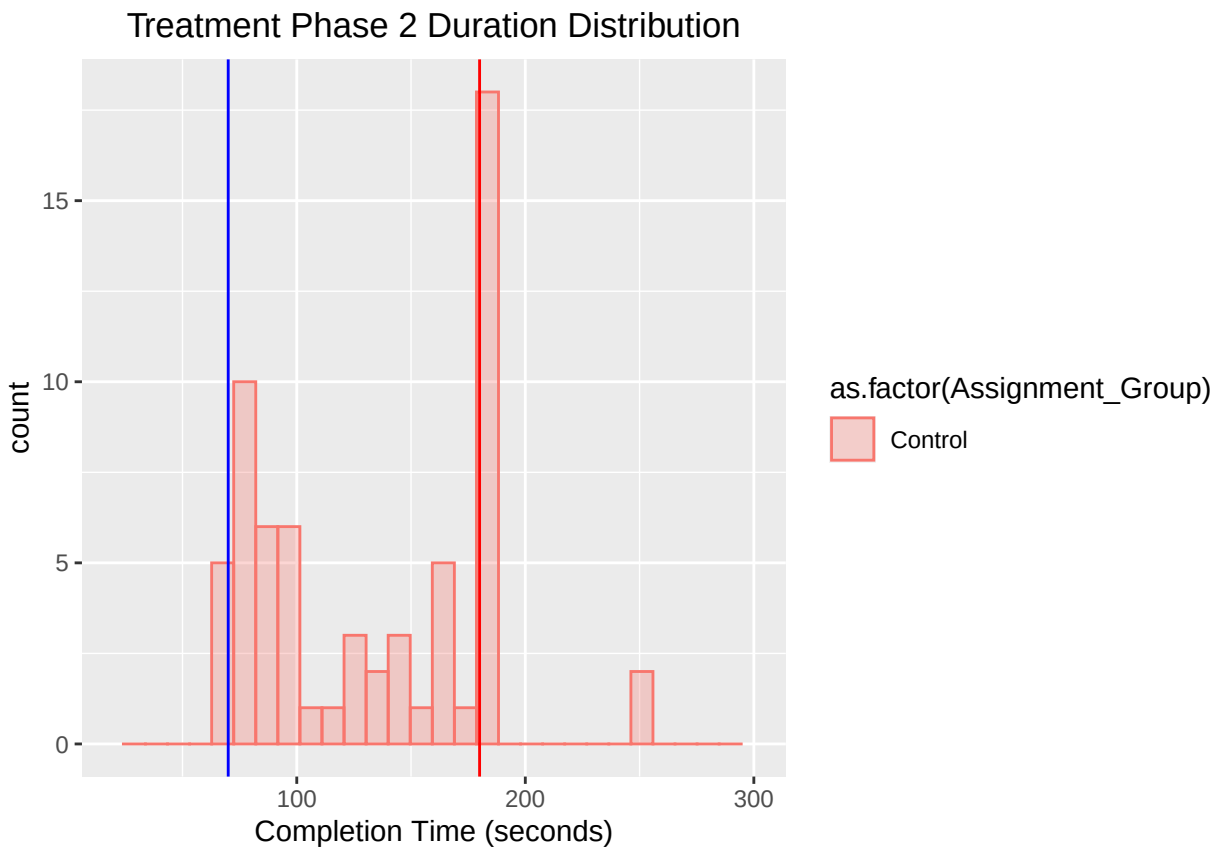
```r
#### set up t-tests comparing the submit times for Treatment Phase I

ggplot(d_noncompliance_1[Assignment_Group %in% c("Control")], aes(x=Treatment_Phase1_SubmitTime, colour=
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 7 rows containing non-finite values (stat_bin).

## Warning: Removed 2 rows containing missing values (geom_bar).
```




```
ggplot(d_noncompliance_2[Assignment_Group %in% c("Control")], aes(x=Treatment_Phase2_SubmitTime, colour=
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 1 rows containing non-finite values (stat_bin).
## Warning: Removed 2 rows containing missing values (geom_bar).
```




```
t.test(d_noncompliance_1[Assignment_Group %in% c("Control"), Treatment_Phase1_SubmitTime],
       d_noncompliance_2[Assignment_Group %in% c("Control"), Treatment_Phase2_SubmitTime])
```

```
##
##  Welch Two Sample t-test
##
## data:  d_noncompliance_1[Assignment_Group %in% c("Control"), Treatment_Phase1_SubmitTime] and d_nonc
## t = 2.4, df = 92, p-value = 0.02
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   5.964 69.955
## sample estimates:
## mean of x mean of y
##     173.5     135.5
```

Throughout this exploratory discussion, we see consistent behavior in which respondents spent less time in the second round of treatment relative to the first. Further investigation will be needed to determine if this behavior is related to the attenuating treatment effects that we observed in our regressions. This also raises questions about whether feedback fatigue triggers noncompliance for our respondents. Nevertheless, when designing our experiment, we expected that people were to be exposed to treatment in both phases at relatively equal times. This discussion illustrates that we should focus on design improvements that encourage people to take treatment consistently.

Power

```
 Two-sample t test power calculation

          n = 161.7
```

```
      delta = 0.05
         sd = 0.16
  sig.level = 0.05
      power = 0.8
alternative = two.sided
```

NOTE: n is number in *each* group

Assignment_Group	N
Negative Images	70
Positive Images	66
Self-Reflect	66
Control	65
Medical Feedback	66

Power analysis shows that our groups did not have a large enough sample size required for each group. Due to the small effect size of approximately 0.05 when comparing mean Task Phase 2 scores in treatment and control groups, for such small effects to be detected with statistical power of 80%, the number of subjects required in each group would be 162. Our group sizes for the control group, as well as the medical, positive, negative, and self-reflect treatment groups were 65, 66, 66, 70, and 66 respectively. This is primarily due to the fact that we charged too high of a price point per completed survey.

Conclusions

Our experiment and following study shows that feedback contributes a statistically and practically significant effect in x-ray analysis performance (ATE = 5.1%, SE +- 2.2%). More specifically, targeted medical feedback saw the largest, statistically significant increases in performance (ATE = 5.6%, SE +- 2.9%), showing that expert opinion may lead to more significant outcomes in the real world. Along the same lines, self-reflection lead to statistically and practically significant improvements on performance (ATE = 5.8%, SE +- 2.9%), which bolsters recent research into the power of self-reflection techniques on a variety of everyday activities. Lastly, negative feedback loops fared the worst (ATE = 3.9%, p = 0.14), showing that for recognition-based tasks, negative feedback may not lead to stronger outcomes than other methods.

Lastly, we found that more frequent feedback loops during a short, iterative task does not lead to significant marginal improvements in performance (ATE = 0.2%, p = 0.92) This may have been due to our experimental design and short duration of the task, but should lead to further research on the relationship between feedback loops and marginal productivity.

This experiment faces potential limitations when making more generalized conclusions about the effects of feedback on performance in addition to lower power. For example, the experiment required analysis of a more simple, X-Ray analysis, which is not as complex of a task when compared to multi-step tasks such as writing a paper or performing 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 scenario. They certainly did not experiment the same time or social pressures or distractions usually present in most constructive feedback instances

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 significant 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), which afforded only 350 participants in our study to comply with the set $500 budget. We should have, however, charged ~ $0.25, which is on par with average MTurk prices per task and would have allowed us to recruit more participants and achieve higher power. These changes would have given the experiment an estimated 2000 participants, with 400 in the control group and in each of the treatment groups. Power for the experiment would have increased substantially and allowed for more meaningful outcomes.

Most notably, our experiment may not generalize well to the external environment because our MTurk worker population may not be a representative cohort of the real working population. In fact, our study participants may reflect more accurately the effect of feedback on people with lower income (income < $150K) and who are younger (age < 50 years old) (Moss & Litman,). In actuality, the MTurk population may benefit the most from feedback because younger people typically have less work experience and may need guidance to further their performance. In addition, people with lower incomes who accept requests through MTurk also demonstrate a desire to improve their financial position, so they may benefit substantially from feedback that drives performance and, subsequently, income (Buchheit et al, 2018).

Survey improvements such as enforcing word limits in specific treatment phases and tracking the time lapsed during task phases would also help us limit noncompliance. We saw several terse responses in the self-reflect and control groups, which are possible signs of noncompliance. Implementing a word limit for respondents in the survey would deter this behavior.

Another improvement that could enhance the experiment would be a factorial design which would allow us to explore heterogeneous treatment effects. Our current implementation explores the differences between four selected types of feedback. However, our analysis would be more robust if we considered whether certain components of feedback were more effective in improving task performance. A factorial design would allow us to consider if it is the presence of a writing component, reading component, or an answer key that improves task performance.