Effect of Learning Feedback Styles on Learning Outcomes

Fall 2020 - W241 Final Report

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

Abstract

Feedback can be used as a useful tool for personal growth and success. While researchers have studied the topic for decades, few controlled studies have been conducted to fully understand the relationship between critique types, feedback loops, and their correlation with successful outcomes. The aim of this study was to assess the effectiveness of several different types of feedback in identifying positive and negative X-Ray images. 350 participants went through an online test session analyzing three sets of X-Ray lung images to determine if they contained pneumonia if they were healthy. Participants were randomly assigned to five different feedback groups and received feedback twice in between the X-Ray imaging sessions.

We found that expert-driven feedback was statistically significant and led to some of the highest improvements in X-Ray analysis. Furthermore, self-reflective feedback techniques were shown to be just as significant and effective. In quick, recognition-based tasks, focusing on negative feedback (i.e. what is wrong) may not be an effective strategy to improve performance. We also found that the marginal improvements in scores from a second feedback session are not significant and may not be worthwhile for shorter duration jobs. Lastly, feedback was found to be more impactful for low achieving performers. High performers do not exhibit any increased boost from feedback and may have been just as successful regardless of feedback sessions.

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 Images 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 Images 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 (225) and India (115). There were more males that participated in the study (207) than females (143).

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.

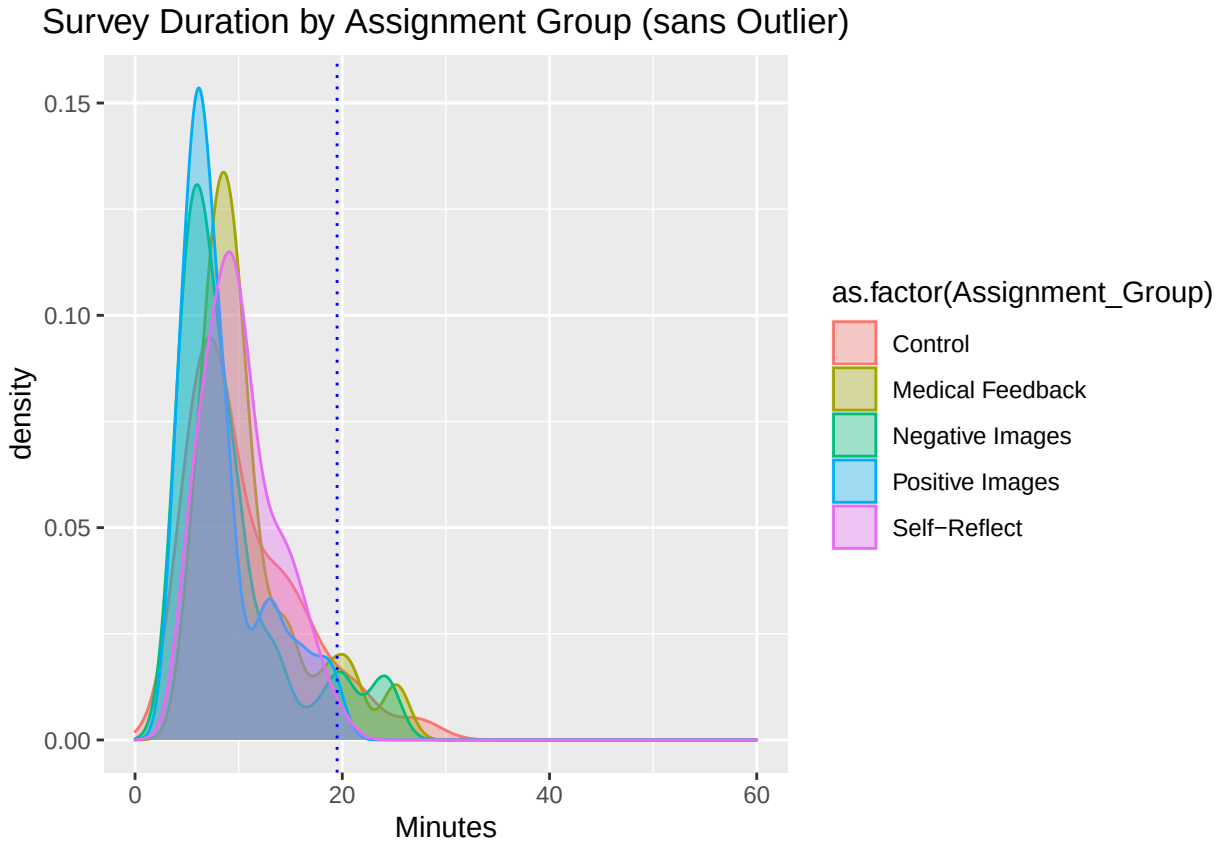
The Qualtrics flow can be seen below.

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.

Figure 1: Qualtrics Flow

```
#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]

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

#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 19.5 minutes or less (6.5 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

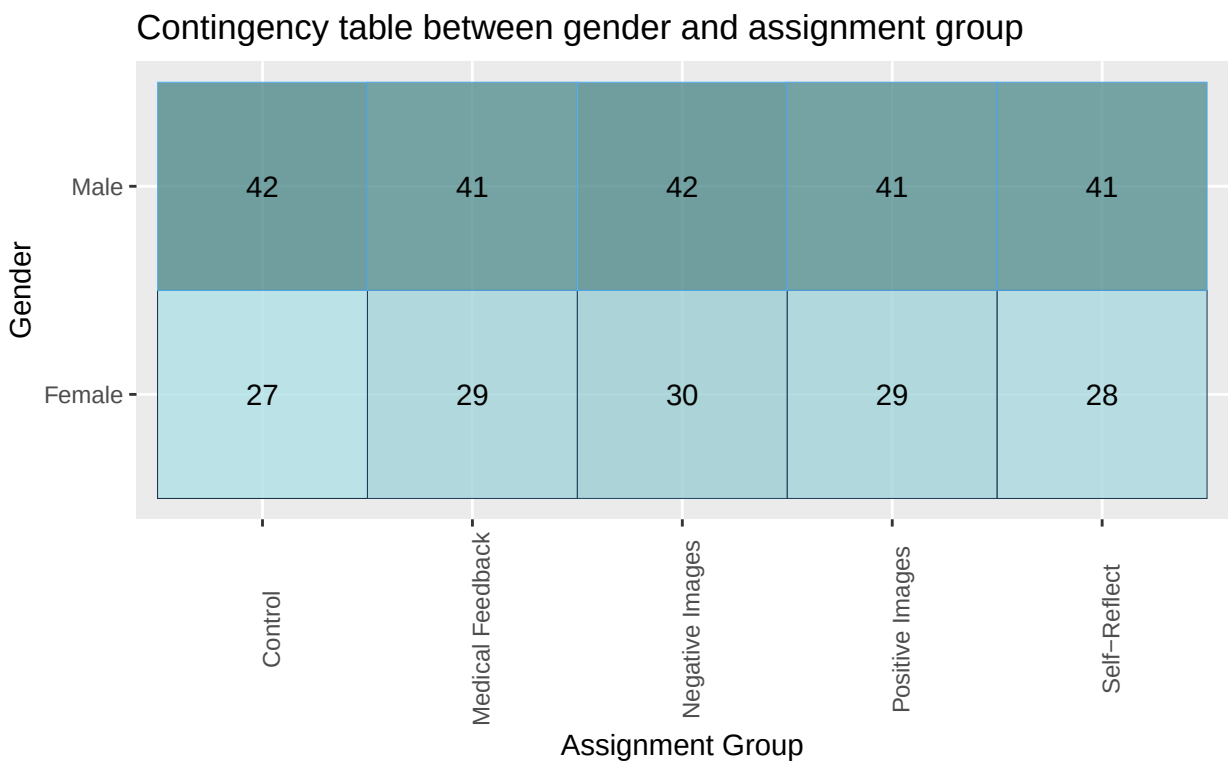
We examined how well our randomization worked by checking that the proportion of individuals assigned to each group was similar. Furthermore, we performed visual covariate balance checks on the survey data as it relates to gender, age range, education, and country. We additionally performed Chi Squared 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 72
## 2:  Positive Images 70
## 3:     Self-Reflect 69
## 4:          Control 69
## 5: Medical Feedback 70
```

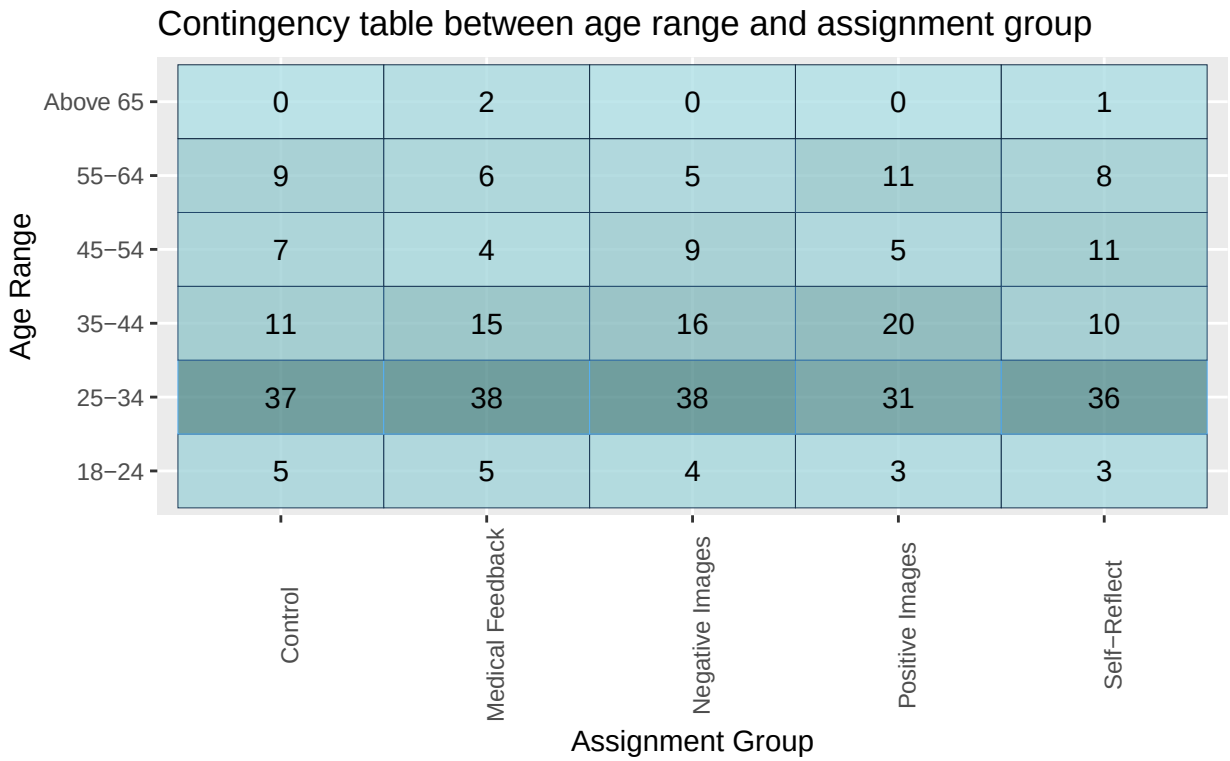
```
# check balance between genders
gender_chisq <- chisq.test(d_respondents[ , table(Assignment_Group, Gender)])
```

```
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))
```



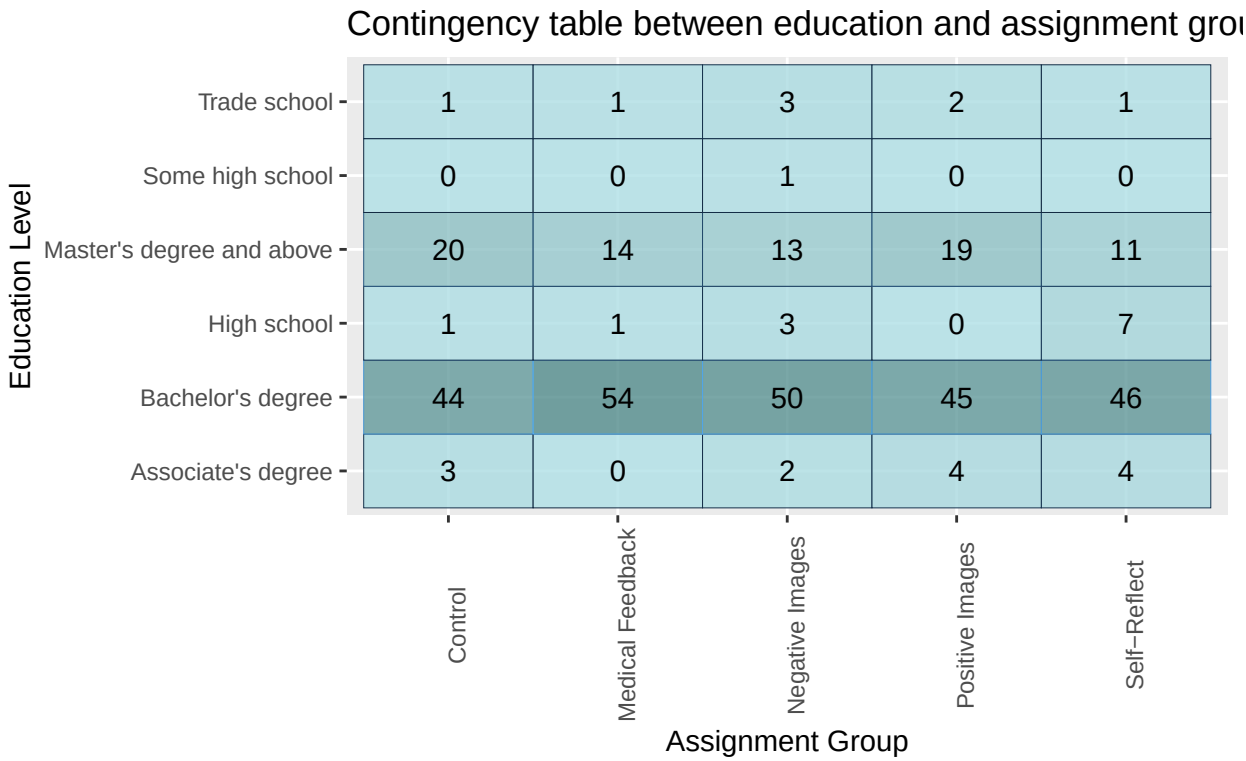
Assuming gender distributions are the same among assignment groups, a chi−squared test for independ
degrees of freedom yields p=0.9981, suggesting that there is no relationship between gender and assign
significance level of 0.05.

```
# 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 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))
```



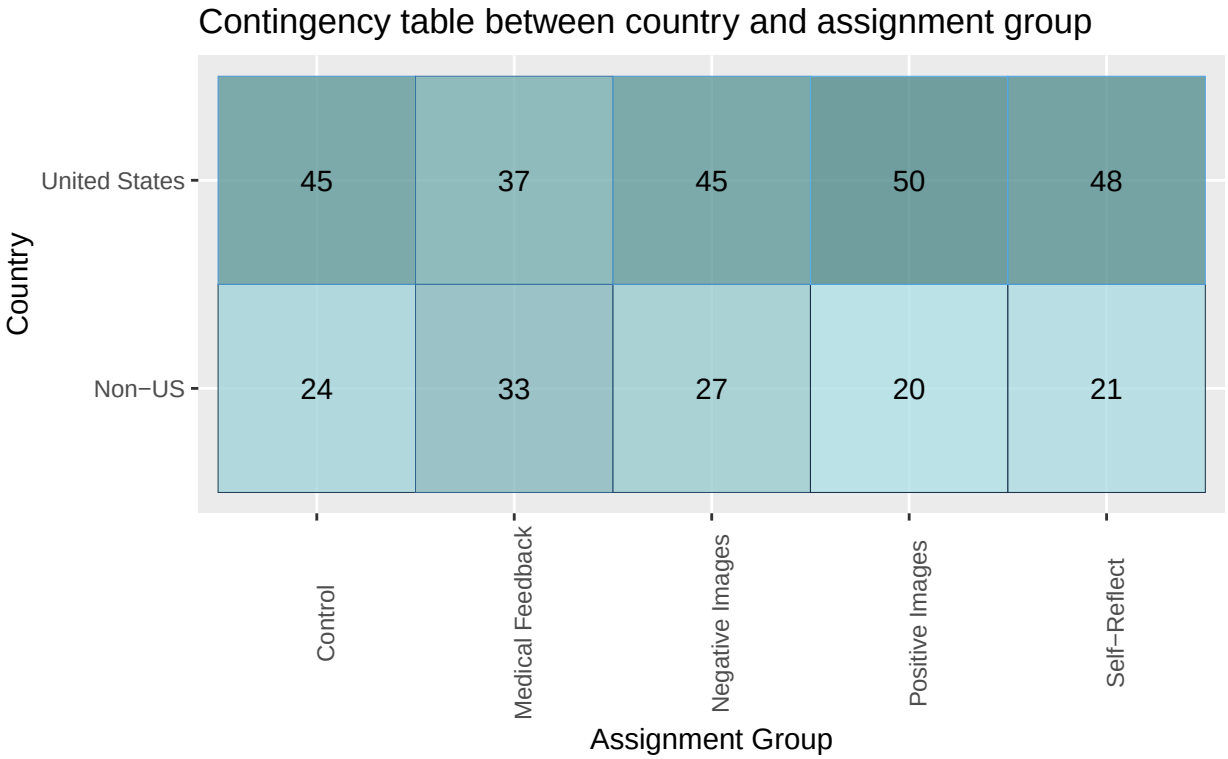
Assuming age distributions are the same among assignment groups, a chi−squared test for independe
Carlo simulation yields p=0.5157, suggesting that there is no relationship between age and assignmen
significance level of 0.05.

```
#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=0.071, 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 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 Specific 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.

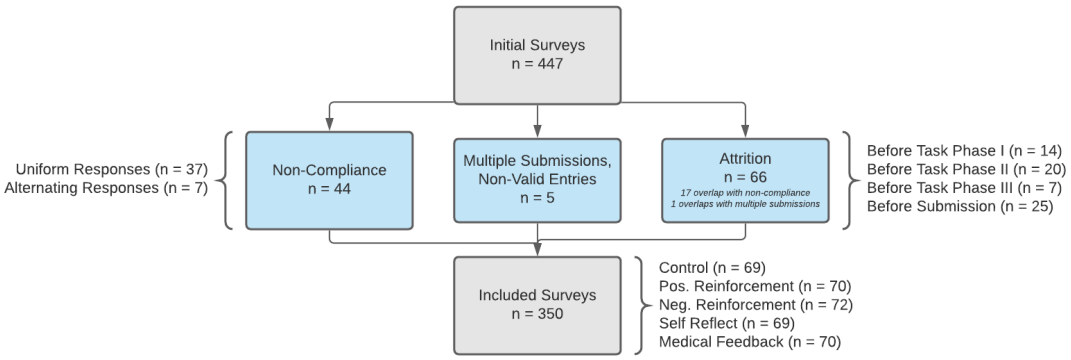
1. Control vs. All Treatment Groups: This compares people who receive the control with people who receive any form of feedback treatment.
2. Individual Treatment Effects: This second comparison focuses on comparing each individual treatment group with the control and with each other.

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:



Our exploratory data analysis digged deeper into the attrition category to see if certain control or feedback groups fell off more than others. By and large the analysis showed little abnormalities between the groups outside of the attrition before Task Phase 1. However, all of the attrition before Task Phase 1 occured in the negative images category. This is likely an design error. However, since this occured ***before*** random assignment, it should not be cause for concern in our analysis.

```
attrition_table <- as.data.frame.matrix(d_attrition[, addmargins(table(Assignment_Group, Attrition_Stag
knitr::kable(attrition_table,
             caption = "Attrition by Stage and Feedback Type",
             footnote = "Random assignment occurs before Task Phase 2")
```

Table 2: Attrition by Stage and Feedback Type

	Before Submission	Before TaskPhase1	Before TaskPhase2	Before TaskPhase3	Sum
Control	4	0	5	2	11
Medical Feedback	5	0	3	1	9
Negative Images	4	14	3	2	23
Positive Images	7	0	2	0	9
Self-Reflect	5	0	7	2	14

	Before Submission	Before TaskPhase1	Before TaskPhase2	Before TaskPhase3	Sum
Sum	25	14	20	7	66

```
attrition_prop_table <- as.data.frame.matrix(addmargins(round(prop.table(d_attrition[, table(Assignment
knitr::kable(attrition_prop_table,
             caption = "Proportion of Attrition by Stage and Feedback Type",
             footnote = "Random assignment occurs before Task Phase 2")
```

Table 3: Proportion of Attrition by Stage and Feedback Type

	Before Submission	Before TaskPhase1	Before TaskPhase2	Before TaskPhase3	Sum
Control	0.06	0.00	0.08	0.03	0.17
Medical Feedback	0.08	0.00	0.05	0.02	0.15
Negative Images	0.06	0.21	0.05	0.03	0.35
Positive Images	0.11	0.00	0.03	0.00	0.14
Self-Reflect	0.08	0.00	0.11	0.03	0.22
Sum	0.39	0.21	0.32	0.11	1.03

```
#control proportion test
prop.test(attrition_table[1,], attrition_table[6,])
```

```
## Warning in prop.test(attrition_table[1, ], attrition_table[6, ]): Chi-squared
## approximation may be incorrect
```

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  attrition_table[1, ] out of attrition_table[6, ]
## X-squared = 4.5, df = 4, p-value = 0.3
## alternative hypothesis: two.sided
## sample estimates:
##         prop 1 prop 2 prop 3 prop 4 prop 5
## Control   0.16      0   0.25 0.2857 0.1667
```

```
#Medical Feedback proportion test
prop.test(attrition_table[2,], attrition_table[6,])
```

```
## Warning in prop.test(attrition_table[2, ], attrition_table[6, ]): Chi-squared
## approximation may be incorrect
```

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  attrition_table[2, ] out of attrition_table[6, ]
## X-squared = 3.1, df = 4, p-value = 0.5
## alternative hypothesis: two.sided
## sample estimates:
##                  prop 1 prop 2 prop 3 prop 4 prop 5
## Medical Feedback    0.2      0   0.15 0.1429 0.1364
```

```r
#Negative Images proportion test
prop.test(attrition_table[3,], attrition_table[6,])
```

```
## Warning in prop.test(attrition_table[3, ], attrition_table[6, ]): Chi-squared
## approximation may be incorrect
```

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  attrition_table[3, ] out of attrition_table[6, ]
## X-squared = 34, df = 4, p-value = 9e-07
## alternative hypothesis: two.sided
## sample estimates:
##                 prop 1 prop 2 prop 3 prop 4 prop 5
## Negative Images   0.16      1   0.15 0.2857 0.3485
```

```r
#Positive Images proportion test
prop.test(attrition_table[4,], attrition_table[6,])
```

```
## Warning in prop.test(attrition_table[4, ], attrition_table[6, ]): Chi-squared
## approximation may be incorrect
```

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  attrition_table[4, ] out of attrition_table[6, ]
## X-squared = 7.9, df = 4, p-value = 0.09
## alternative hypothesis: two.sided
## sample estimates:
##                 prop 1 prop 2 prop 3 prop 4 prop 5
## Positive Images   0.28      0    0.1      0 0.1364
```

```r
#Self Reflect proportion test
prop.test(attrition_table[5,], attrition_table[6,])
```

```
## Warning in prop.test(attrition_table[5, ], attrition_table[6, ]): Chi-squared
## approximation may be incorrect
```

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  attrition_table[5, ] out of attrition_table[6, ]
## X-squared = 6.3, df = 4, p-value = 0.2
## alternative hypothesis: two.sided
## sample estimates:
##              prop 1 prop 2 prop 3 prop 4 prop 5
## Self-Reflect    0.2      0   0.35 0.2857 0.2121
```

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

ımpare task scores in different phases

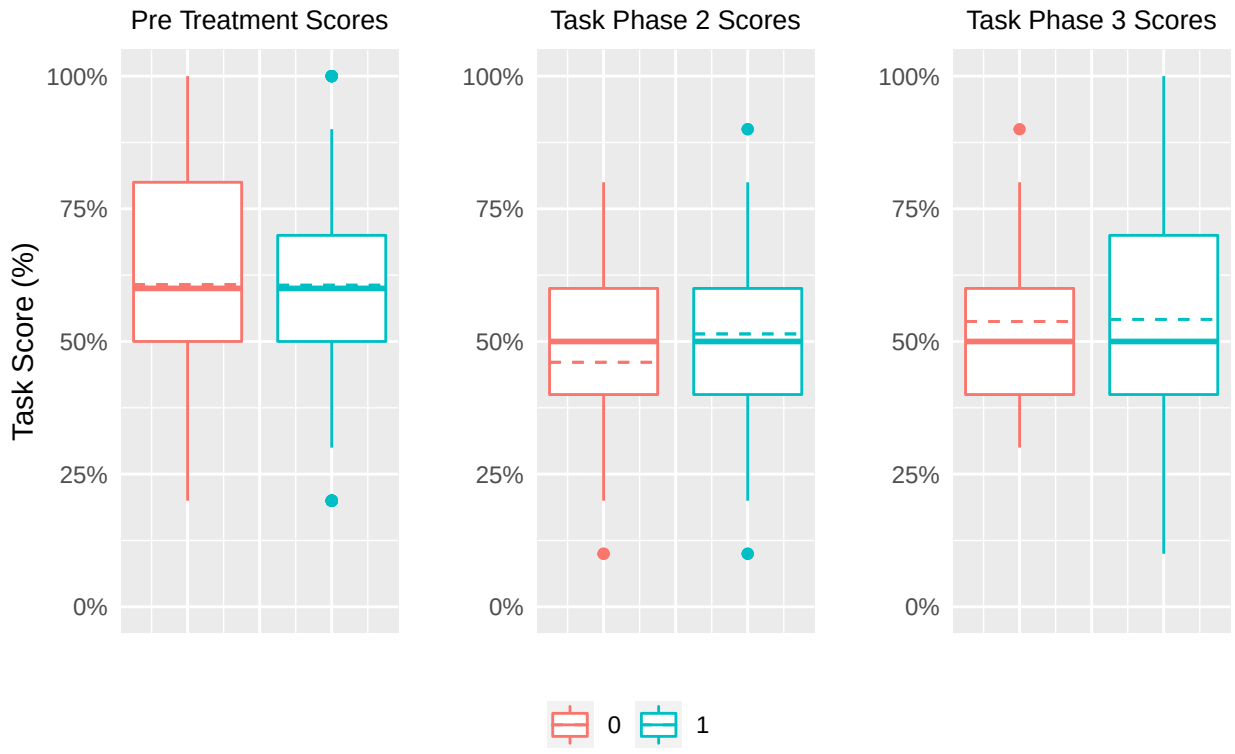
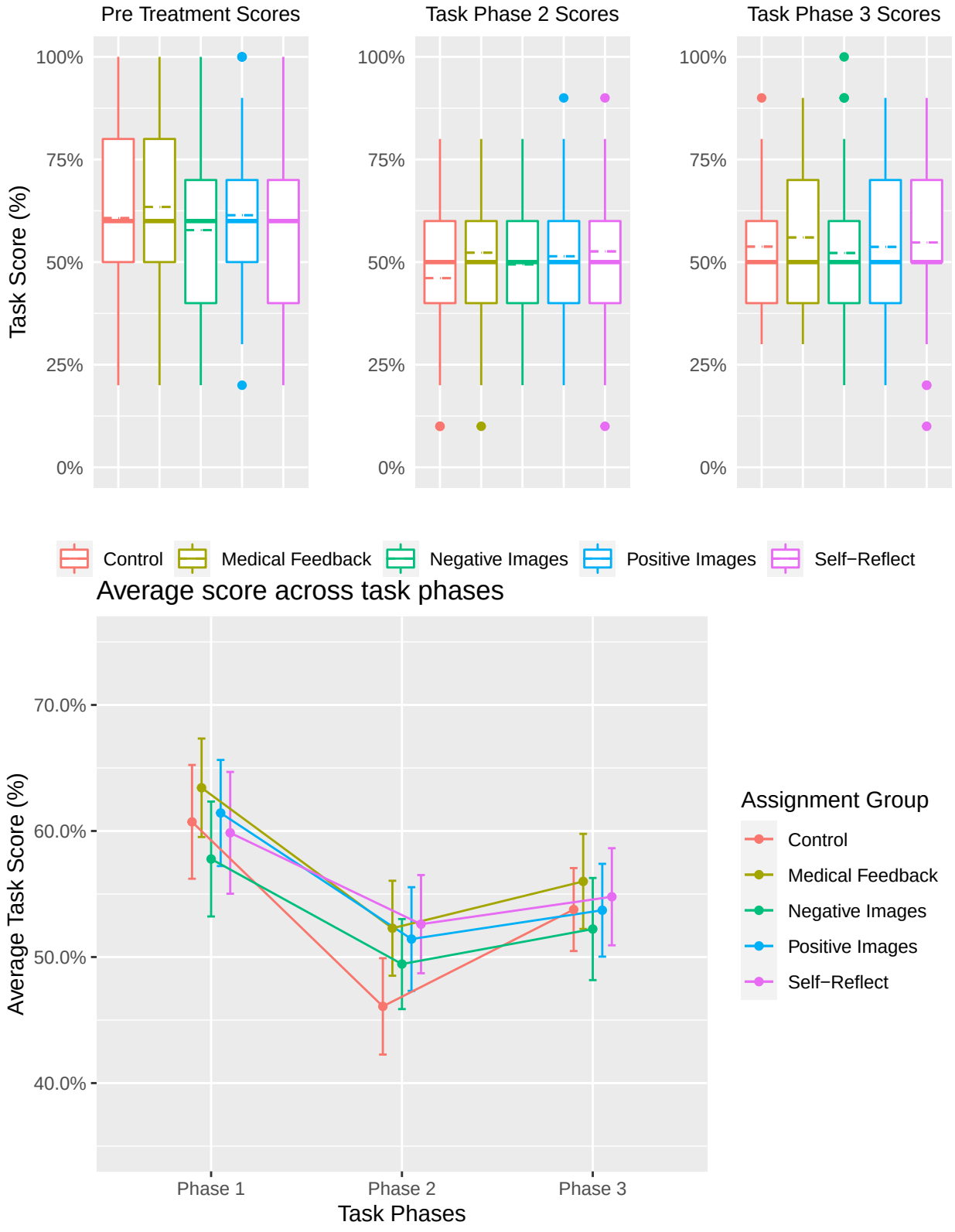


Table 4: Welch Two Sample t-test: `d_respondents[Treatment_Dummy == 0, TaskPhase1_Score]` and `d_respondents[Treatment_Dummy == 1, TaskPhase1_Score]`

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
0.04674	102.6	0.9628	two.sided	0.6072	0.606

ɔmpare task scores in different phases



Regressions

Table 5:

	Dependent variable:			
	Task Phase 2 Score			
	(1)	(2)	(3)	(4)
Any Treatment	0.053**	0.051**		
	(0.022)	(0.022)		
Medical Feedback			0.062**	0.056*
			(0.027)	(0.029)
Negative Images			0.034	0.040
			(0.027)	(0.027)
Positive Images			0.053*	0.050*
			(0.029)	(0.028)
Self-Reflection			0.065**	0.059**
			(0.028)	(0.029)
Task Phase 1 Score		0.241***		0.238***
		(0.047)		(0.048)
Male		−0.009		−0.009
		(0.018)		(0.018)
US		0.007		0.008
		(0.021)		(0.022)
Constant	0.461***	0.277***	0.461***	0.279***
	(0.019)	(0.073)	(0.019)	(0.074)
Education FE	No	Yes	No	Yes
Age FE	No	Yes	No	Yes
Observations	350	350	350	350
R^2	0.017	0.118	0.021	0.119
Adjusted R^2	0.014	0.081	0.010	0.074
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
335	8.335	NA	NA	NA	NA
332	8.32	3	0.01466	0.195	0.8998

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 feedback from domain expertise 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 specific 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 8: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
335	8.116	NA	NA	NA	NA
332	8.098	3	0.01802	0.2463	0.8639

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 7:

	Dependent variable:			
	Task Phase 3 Score			
	(1)	(2)	(3)	(4)
Any Treatment	0.004	0.002		
	(0.019)	(0.019)		
Medical Feedback			0.022	0.011
			(0.026)	(0.026)
Negative Images			−0.015	−0.011
			(0.027)	(0.026)
Positive Images			−0.001	0.004
			(0.025)	(0.025)
Self-Reflection			0.010	0.005
			(0.026)	(0.026)
Task Phase 1 Score		0.161***		0.157***
		(0.047)		(0.047)
Male		−0.004		−0.004
		(0.017)		(0.017)
US		−0.005		−0.004
		(0.020)		(0.020)
Constant	0.538***	0.518***	0.538***	0.520***
	(0.017)	(0.065)	(0.017)	(0.064)
Education FE	No	Yes	No	Yes
Age FE	No	Yes	No	Yes
Observations	350	350	350	350
R^2	0.0001	0.085	0.006	0.087
Adjusted R^2	−0.003	0.046	−0.005	0.040
Note:				*p<0.1; **p<0.05; ***p<0.01

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	72
Positive Images	70
Self-Reflect	69
Control	69
Medical Feedback	70

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 targeted medical feedback, 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

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 most 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 = 5.9%, 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 = 4.0%, 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, which 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 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).