Effect of Learning Feedback Styles on Learning Outcomes

Fall 2020 - W241 Final Report

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

Contents

1 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.

2 Background

Whether its the coach and player, teacher and pupil, or managers and direct reports, feedback is used pervasively in organizations with the objective of driving performance improvement at the task at hand. All leaders are encouraged to give feedback while understudies are taught to receive it openly. However, what is good feedback and how much of one's success on a given task be attributed to this feedback? A large number of studies looked into the relationship between feedback and performance, summarized in a meta-analysis by Kluger & DeNisi (1996) (The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin, 119(2), 254–284). Kluger & DeNisi review shows that while "feedback interventions" overall incrased performance, they in some cases led to decreased performance. In this study, we seek to better understand how feedback influences successful outcomes and whether different types of feedback lead to better outcomes than others.

3 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 can take various forms and be delivered as advice, constructive critique, or in some cases as harsh criticism. 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.

3.1 Task selection

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). It also allows to measure outcome on an objective scale (number of images classified correctly) with significant data available on classified images.

3.2 Hypothesis

Our study seeks to answer the following question:

> *What type of feedback (detailed feedback, 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.

4 Experimental Design

4.1 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 (after the first round of treatment) in order to test the immediate effects of feedback on performance. We further compare the first iteration scores with those in the third iteration (after the second round of treatment) 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 a situation where a manager provides one dimension of feedback to drive pattern recognition - in this case only images that should have passed the test.
- *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 a situation where a manager provides one dimension of feedback to drive pattern recognition - in this case only images that should not have passed the test.
- *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.

4.2 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, 2020	Nov. 6 - 8,2020	Nov. 9 - 14, 2020	Nov. 15 - 30, 2020	Dec. 8, 2020	Dec. 15, 2020

4.3 Enrollment and Recruitment Process

Subjects were recruited through Mechanical Turk (MTurk) and received USD 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).

4.4 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.

4.5 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.

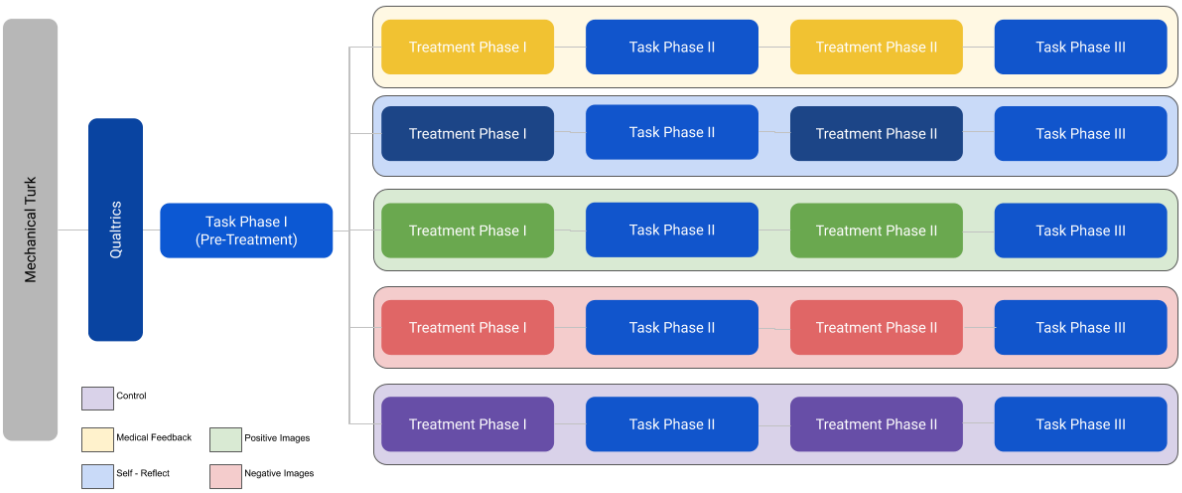
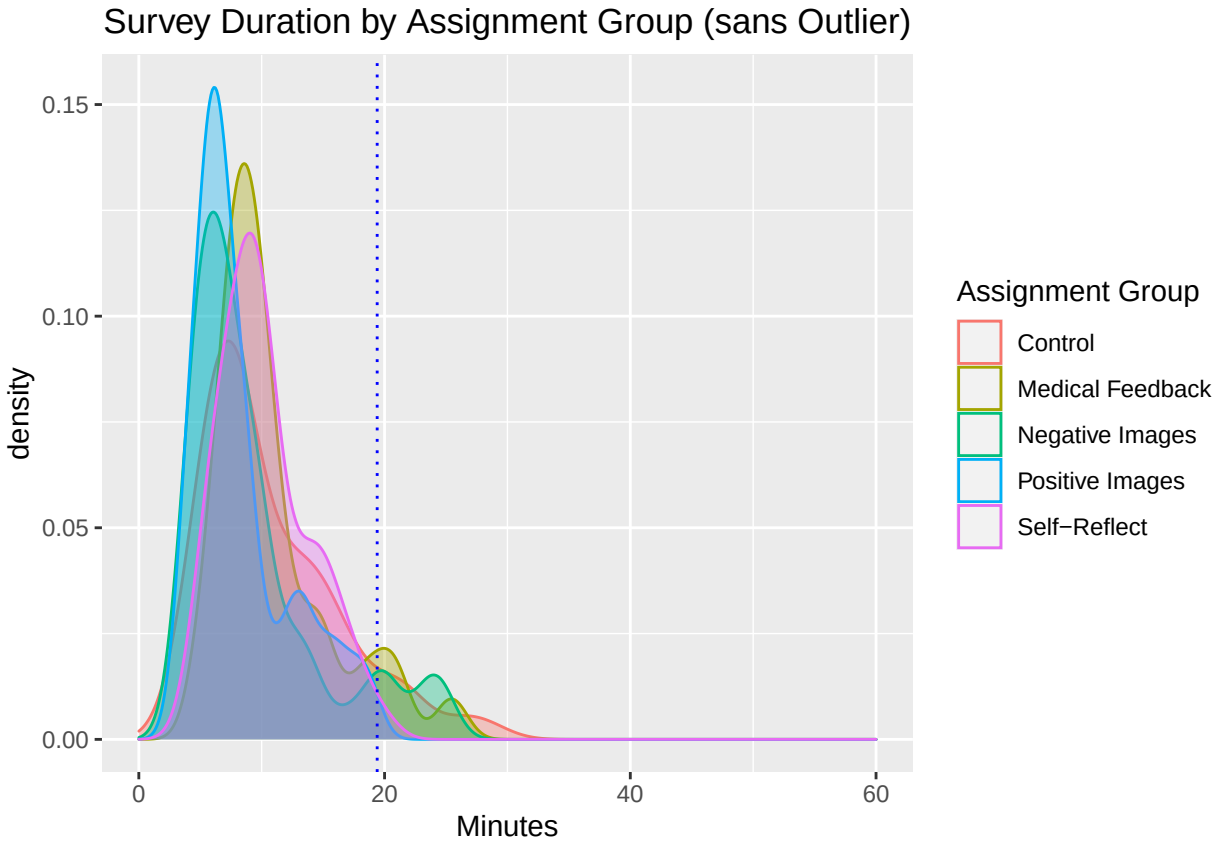


Figure 1: Qualtrics Flow

4.6 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.



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 (6.467 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.

4.7 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.

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

`Chi-squared test for given probabilities`

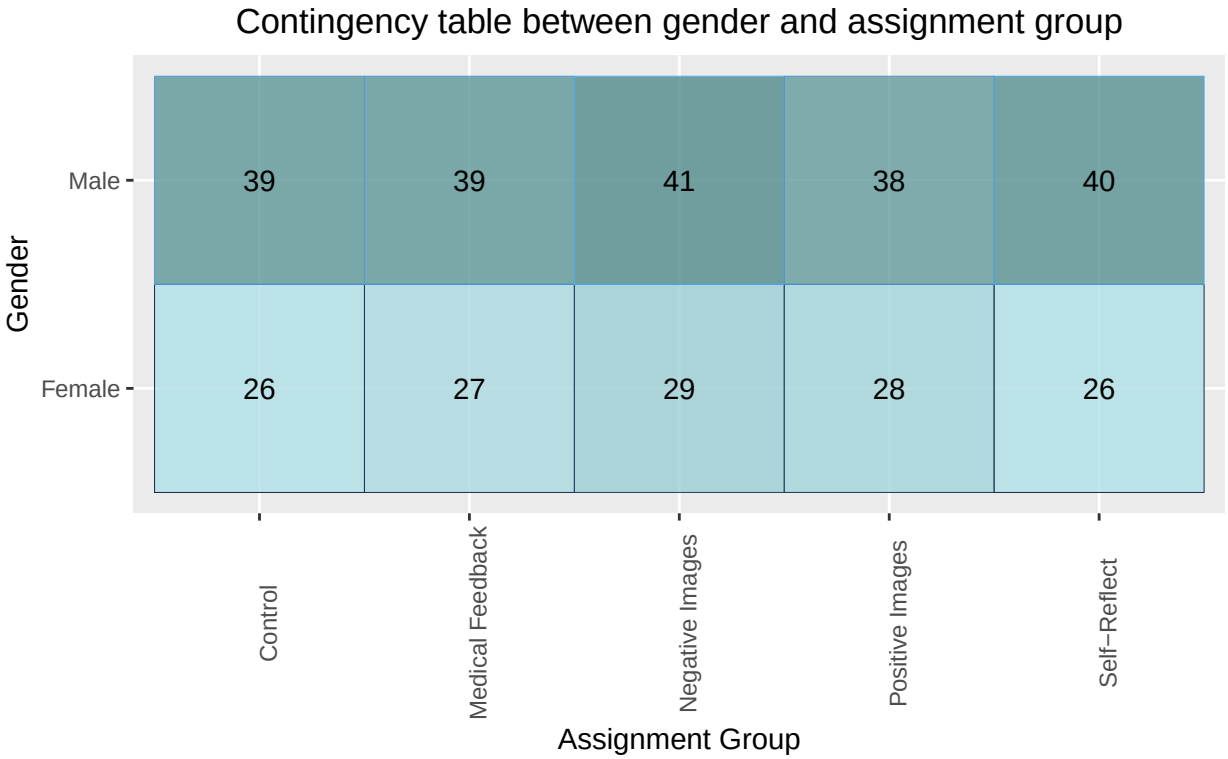
data: d_respondents[, table(Assignment_Group)] X-squared = 0.2, df = 4, p-value = 1

Table 3: Welch Two Sample t-test: `d_respondents[Treatment_Dummy == 0, TaskPhase1_Score]` and `d_respondents[Treatment_Dummy == 1, TaskPhase1_Score]`

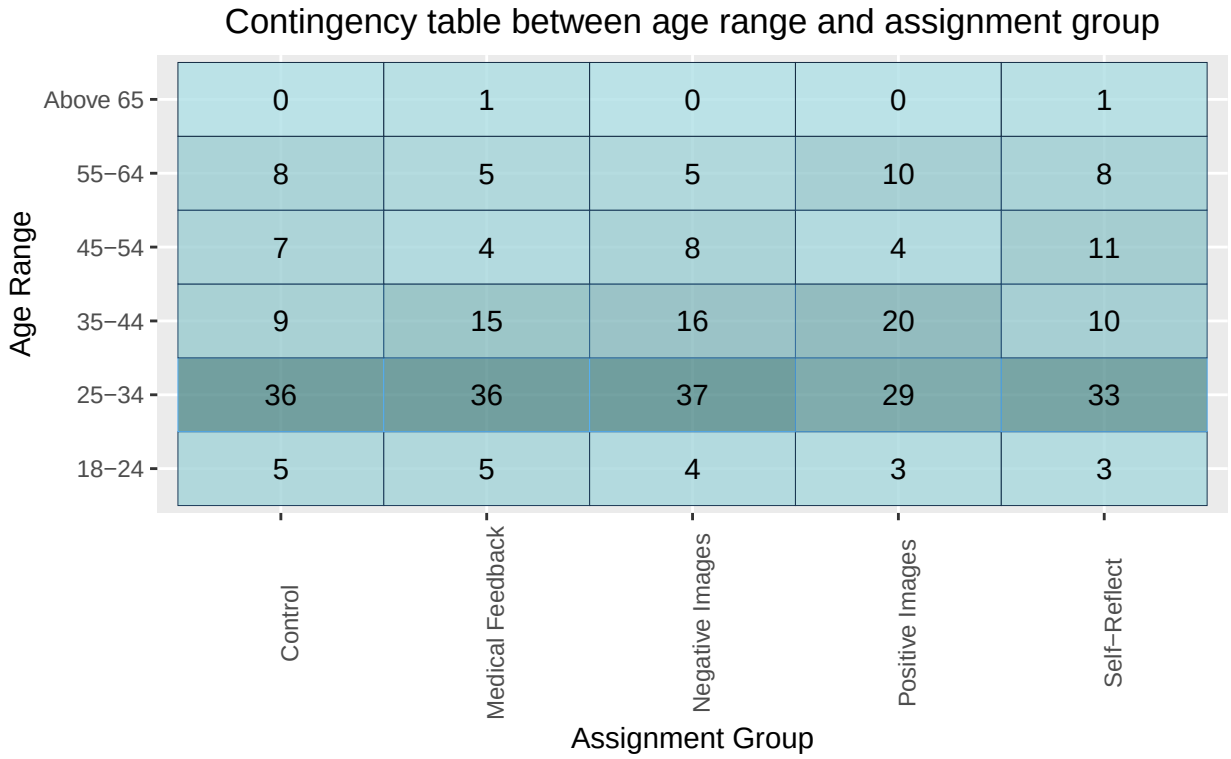
Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
0.07098	95.68	0.9436	two.sided	0.6138	0.6119

Table 4: Analysis of Variance Table

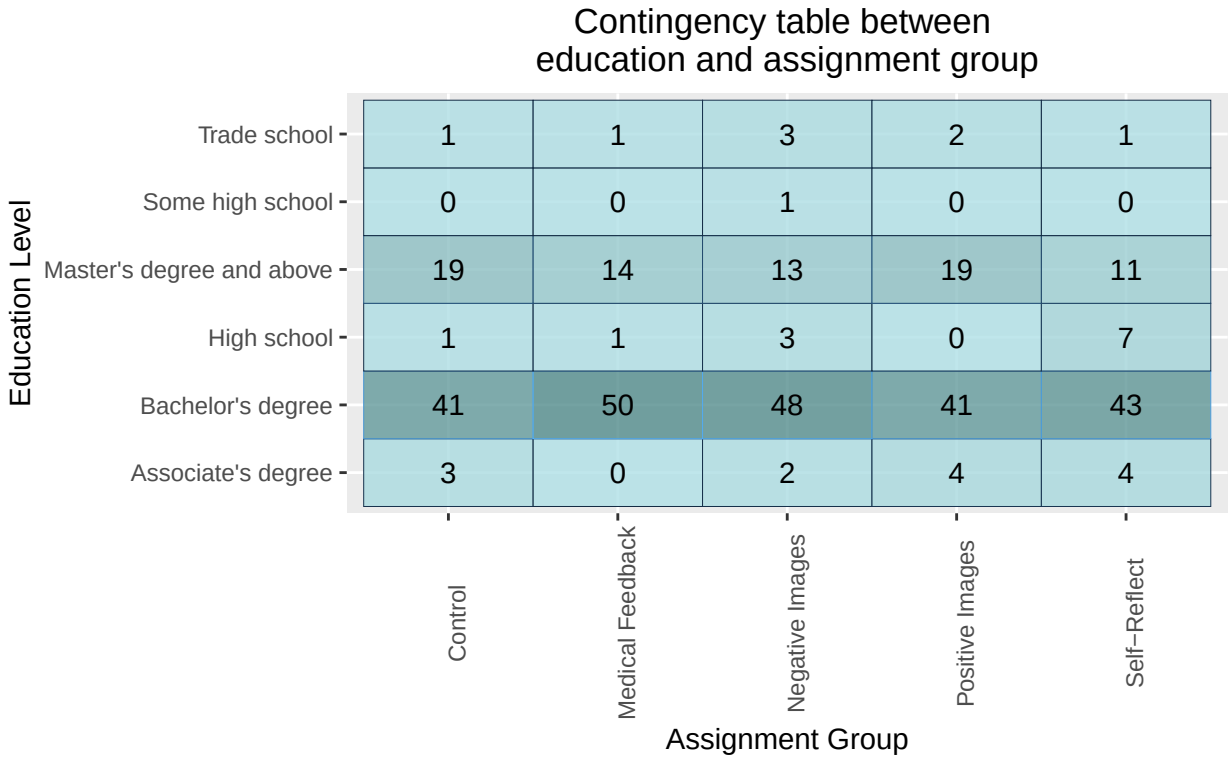
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Assignment_Group	4	0.1485	0.03712	1.019	0.3977
Residuals	328	11.95	0.03644	NA	NA



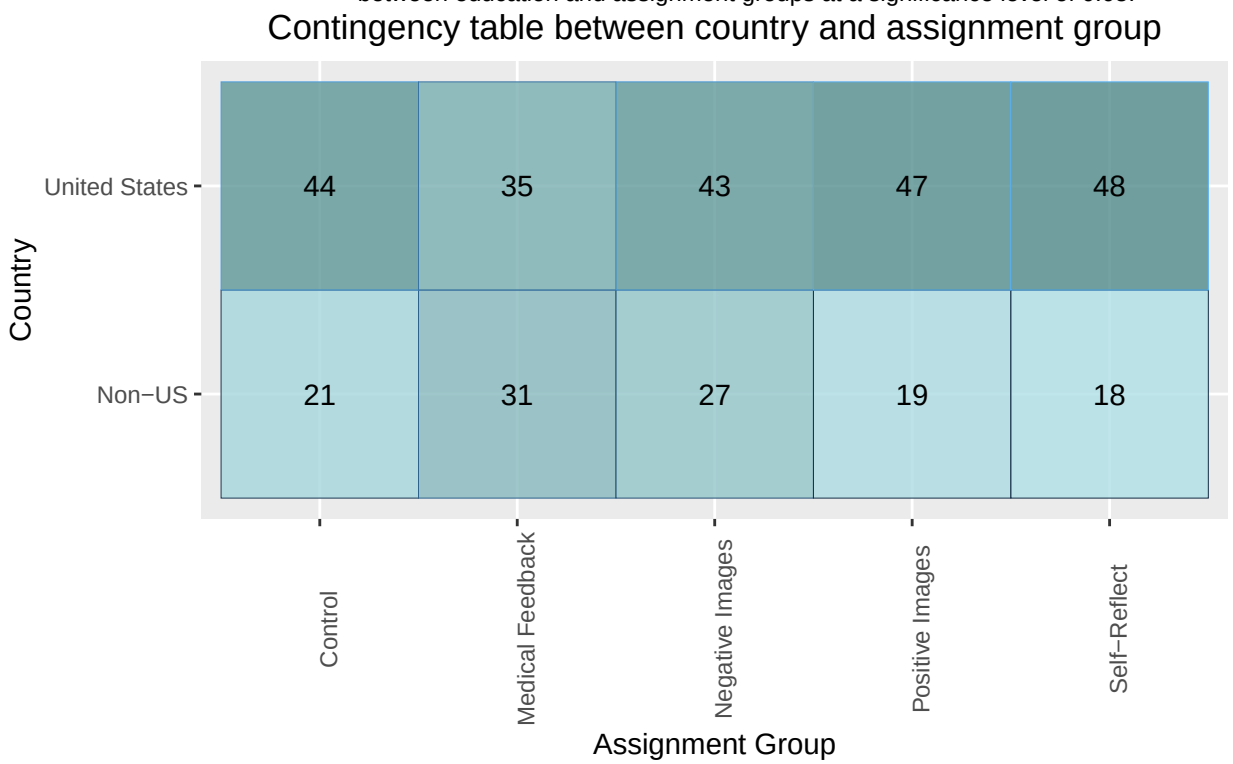
Assuming gender distributions are the same among assignment groups, a chi−squared test for independ degrees of freedom yields p=0.9972, suggesting that there is no relationship between gender and assign significance level of 0.05.



Assuming age distributions are the same among assignment groups, a chi−squared test for independe Carlo simulation yields p=0.5212, suggesting that there is no relationship between age and assignmen significance level of 0.05.



Assuming education distributions are the same among assignment groups, a chi−squ
independence with Monte Carlo simulation yields p=0.0825, suggesting that there is
between education and assignment groups at a significance level of 0.05.



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.

4.8 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)

Scoring is based on the number of questions a person answers correctly out of 10 questions per phase, which is then converted to a ratio value for ease of interpretation. In this case, a 10 percentage point increase in performance would signify getting 1 additional question right.

We will assess two main regressions with the following outcome variables: Task Phase 2 Scores and Task Phase 3 Scores. In the former regression, we assess whether feedback immediately affects performance; in the latter analysis, we asses whether there is a marginal increase in performance from repeated feedback.

Within this problem space, 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.

4.9 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:

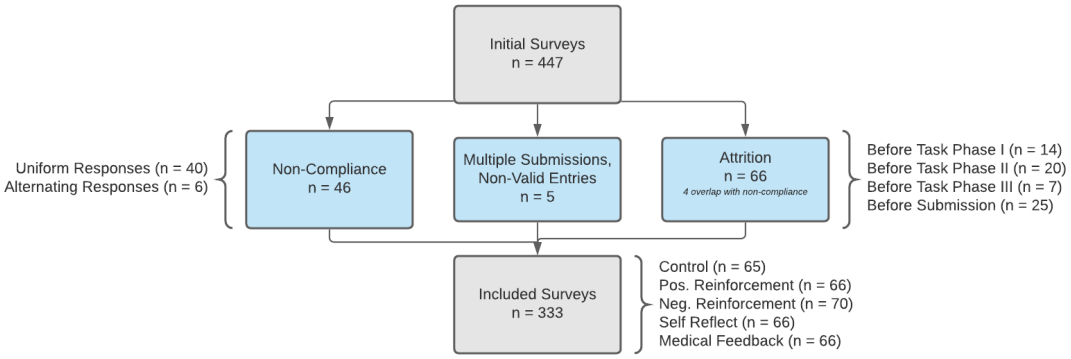
1. Clear non-compliance (n = 46): Some participants did not give honest effort on the survey and answered all "Normal", all "Pneumonia", or all alternating responses during Task Phase 1. This phase served as both a method to garner pre-treatment scores as well as screen for non-compliers. As such, results that aligned with this definition were treated as instances of non-compliance and thrown out of the survey; this methodology would not bias our estimates because we conducted a placebo design to assess the number of Never Takers that occur in our control/placebo group.
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:

Table 5: Attrition by Stage and Feedback Type

	Before TaskPhase2	Before TaskPhase3	Before Submission	Sum
Control	5	2	4	**11**
Medical Feedback	3	1	5	**9**
Negative Images	3	2	4	**23**
Positive Images	2	0	7	**9**
Self-Reflect	7	2	5	**14**
Sum	**20**	**7**	**25**	**66**

Note:
Random assignment occurs before Task Phase 2



Our exploratory data analysis digged deeper into the attrition category to see if certain control or feedback groups fell off more than others. Since assignment occurred after Task Phase 1, our survey design did not allow for us to track subject's attrition by grouping before TaskPhase 1. Qualtrics automatically assigned all attrition pre-Task Phase 1 to the negative images group. Each instance of attrition before TaskPhase 1 was falsely masked in the Negative Images category, skewing any post-assignment analysis. As such we were unable to breakout attrition by treatment group before assignment.

As a next step, we compared attrition rates within treatment groups and within phases as described in Table 5. We found that attrition rates were not statistically significant within these comparisons using a Fisher test ($p>>.05$). Since attrition is not due to any one task phase stage or treatment group, we do not have much reason for concern going forward and can contribute this attrition to randomness rather than systematic effects from any particular treatment or the experimental design.

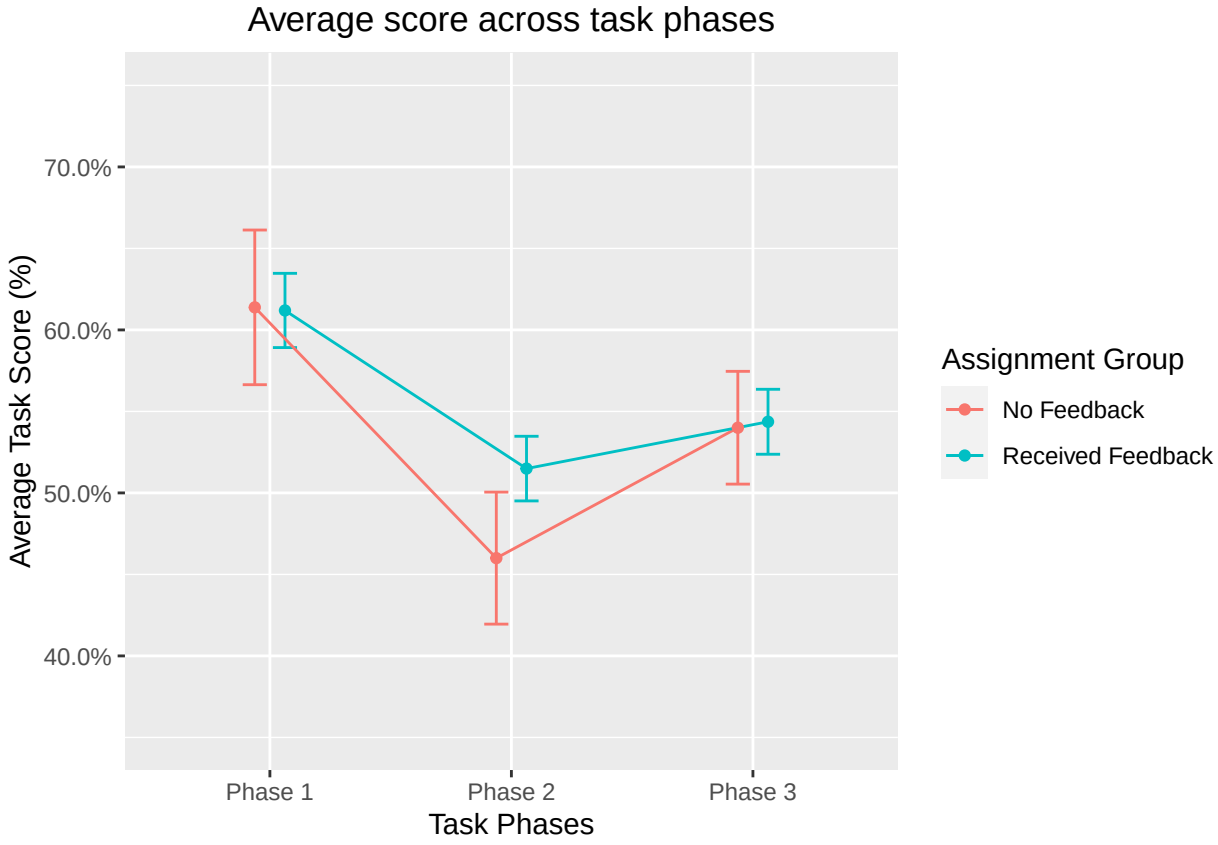
5 Results

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.

5.1 Overview

5.1.1 Immediate Effects of Feedback



When comparing task scores for across people who received feedback and people who received the placebo (shown above), we see that in Task Phase 1, average task score percentage is fairly similar between groups ($\bar{x}_{treatment}$ = 0.612 (SE=0.023), $\bar{x}_{control}$ = 0.614 (SE=0.047)). As shown in the Covariate Balance Checks Section, specifically Table 3, there is no statistical significance between pre-treatment scores in the binary assignment group case (p=0.944).

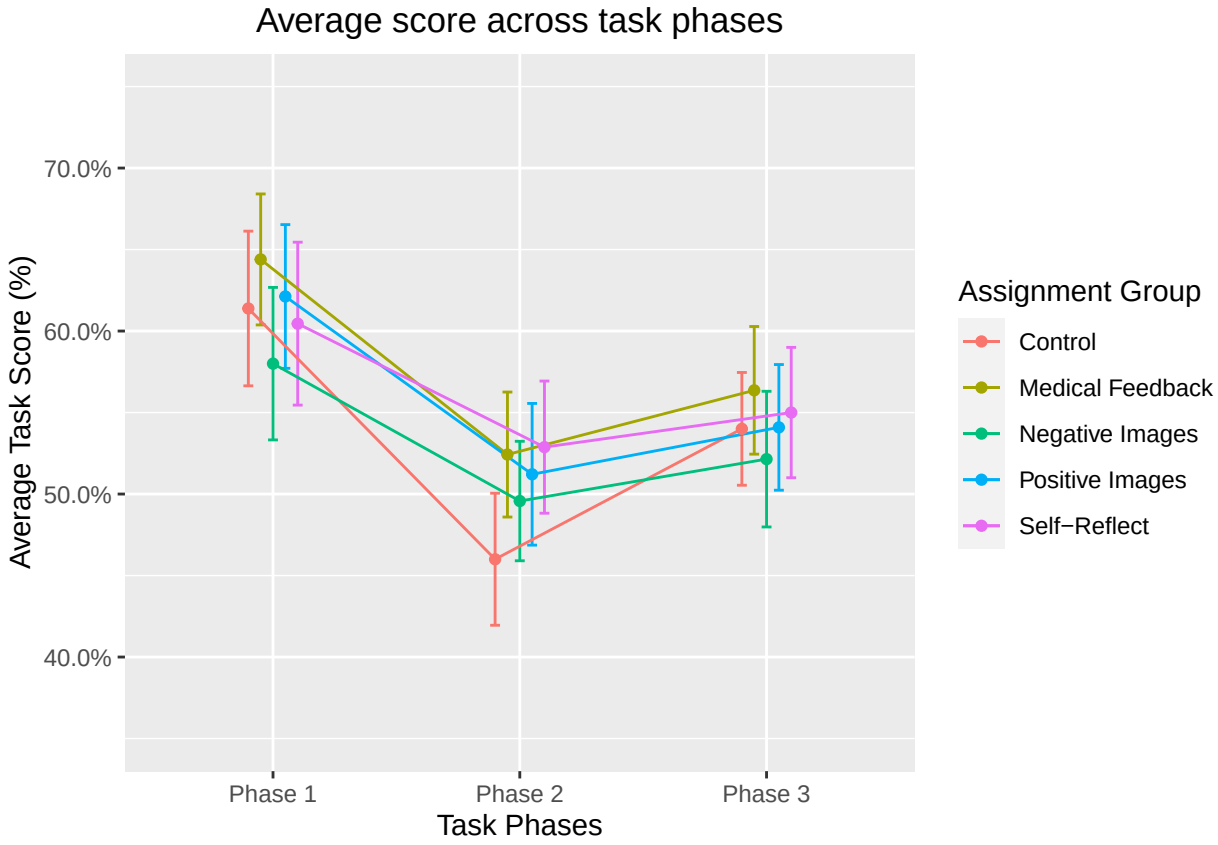
In Task Phase 2, there is a notable difference in performance after the treatment group received feedback and the control group received the placebo, which a t-test in Table 6 deems statistically significant (p=0.019). Furthermore, there is an overall drop in performance between Phase 1 and Phase 2, which suggests that the Phase 2 scores were more difficult compared to those in Phase 1.

Task Phase 3 scores recovered across both binary assignment groups (t(110.615) = -0.18, p=0.858), which may indicate that more rounds of feedback within this timespan do not makea significant impact compared with the placebo (see Effects of Repeated Feedback for more information).

Table 6: Welch Two Sample t-test: `d_respondents[Treatment_Dummy == 0, TaskPhase2_Score]` and `d_respondents[Treatment_Dummy == 1, TaskPhase2_Score]`

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
-2.388	97.01	0.0189 *	two.sided	0.46	0.5149

5.1.2 Effects of Repeated Feedback



Repeating the former analysis on the individual treatment group parses out any substantive differences in average score across phases. Reviewing the average score within Task Phase 1 for individual treatment groups suggests there is no difference between various types of feedback (p=0.398).

Most notably, respondents in the Medical Feedback Group typically score higher than the rest of the survey pool, whereas respondents in the Negative Images Group typically score lower than people in other feedback groups. However, as shown in the figure above, there is much overlap in 95% confidence intervals within each task phase, suggesting that any difference is not statistically significant at the 5% significance level.

5.2 Regressions

5.2.1 Immediate Effects of Feedback

In order to assess the immediate effects of receiving feedback, we consider the outcome measure, `Task Phase 2 Score`, which is a post-treatment variable that measures the effect of one round of feedback/placebo (Table 7).

In columns 1 and 2 of Table 7, we assess the combined effect of all feedback treatment groups by creating a treatment dummy variable `Any Feedback`. We therefore simulate the real world phenomenon where managers have diverse ways of giving feedback, but the direct reports still receive some semblance of a performance review.

Table 8: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
331	9.096	NA	NA	NA	NA

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
318	8.147	13	0.9495	2.851	0.0006707

Column 2 of this table includes the effect of the covariates (pre-treatment score, gender, and FE from education, age, and country), which is shown to be necessary according to F-test results in Table 8. With all variables remaining constant, subjects experience a 5.307 (2.352) percentage point increase in task performance when any feedback is given to the survey respondents, which is statistically significant given an $\alpha = 0.05$. In a notable discovery, adding in these covariates does not lead to a marginal decrease in standard error, so the ATE is no more precise when controlling for these other variables.

To confirm pre-existing assumtpions around high-performers, we find that each 10 percentage point increase in Task Phase 1 scores is associated with a 2.409 (SE=0.476) percentage point increase in performance in Task Phase 2; people who perform well before feedback may also perform well after feedback because high-performers do not require the additional feedback to be successful. This finding naturally motivated an analysis around heterogeneous treatment effects between feedback and high-performers, which found that a 10 percentage point increase in Task 1 Phase score yields an increase of 0.3 (SE=0.518) percentage points, but this is not statistically significant, suggesting that there is no incremental benefit of treatment for high performers (Column 3).

Based on these findings, we then explored the type of feedback that would yield the most positive impact on task performance (Column 4). In doing so, we aimed to inform managers the type of feedback that would garner better performances from their direct reports. We theorized that feedback from domain expertise would foster the highest ATE 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 individual success. At a high level, we see that when people receive specific medical feedback, they experience a 5.882 (SE=2.999) percentage point increase in performance that is statistically significant only at the $\alpha = 0.1$ level, which suggests we require more research to improve the precision around this estimate.

We hypothesized that the negative images feedback would fare the worst because only results from the pneumonia images are shared. In doing so, we simulated when a manager focuses on giving feedback only in abnormal situations rather than promoting standards met day-to-day. As a result, direct reports may have a poorer understanding of what "normal" or "good" looks like. **TODO INSERT STUDY FROM DAHLER** In Column 4, people in the negative image feedback group have only a 4.355 (SE=2.779) percentage point increase that is not statistically significant, indicating that negative feedback was not helpful in improving performance.

Surprisingly, people who were asked to self-reflect on their responses had a statistically significant 6.132 (SE=3.068) percentage point increase in performance. This type of feedback is particularly interesting because self-reflection is a common personal growth technique that is touted in articles in HBR, Forbes, etc. Through this study, we were able to confirm the positive effects of self-reflection; as a manager, you might encourage this behavior through incorporating self-assessments, although in the [Generalizeability Section][Generalizeability], we discuss the caveats of applying these findings outside of the lab environment.

Table 9: 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

Lastly, an F-test shown in Table 9 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 (p=0.915).

Table 7: Test for immediate effects of feedback on performance

	Dependent variable:			
	Task Phase 2 Score			
	(1)	(2)	(3)	(4)
Any Feedback	0.055**	0.053**	0.051**	
	(0.023)	(0.024)	(0.026)	
Medical Feedback				0.059**
				(0.030)
Negative Images				0.044
				(0.028)
Positive Images				0.049*
				(0.029)
Self-Reflect				0.061**
				(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)
High Performer			0.079*	
			(0.047)	
Any Feedback:High Performer			0.030	
			(0.052)	
Constant	0.460***	0.276***	0.437***	0.277***
	(0.021)	(0.074)	(0.023)	(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	0.090	0.121
Adjusted R^2	0.014	0.081	0.081	0.074
Note:				*p<0.1; **p<0.05; ***p<0.01

5.2.2 Effects of Repeated Feedback

Table 10: Test for effects of repeated feedback on performance

	Dependent variable:		
	Task Phase 3 Score		
	(1)	(2)	(3)
Any Feedback	0.004	0.004	
	(0.020)	(0.020)	
Medical Feedback			0.015
			(0.027)
Negative Images			−0.013
			(0.027)
Positive Images			0.007
			(0.026)
Self-Reflect			0.006
			(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***	0.524***
	(0.018)	(0.065)	(0.065)
Education FE	No	Yes	Yes
Age FE	No	Yes	Yes
Observations	333	333	333
R^2	0.0001	0.082	0.085
Adjusted R^2	−0.003	0.042	0.036
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 11: 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

We test the effect of multiple rounds of feedback on task performance using Task Phase 3 as an outcome variable since this task phase occurs after the subjects have received two rounds of treatment or placebo. We anticipate that more feedback received will yield even higher task performance scores compared to Task Phase 2. As a result, we would like to assess if, as a manager, he/she should instantiate more touchbases to review performance.

However, according to Table 10, 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 (Column 2), there is a meager 0.366 (SE=2.031) percentage point increase in performance, which is not statistically significant. Furthermore, expanding the analysis to illustrate individual treatment effects suggests that the varying feedback types do not garner improved performance after more rounds of feedback (Column 3). Indeed, an F-test shown in Table 11 indicates that the expanded model does not add significant value to data

representation (p=0.772).

The findings may be attributed to a number of underlying factors. For example, more frequent feedback during this short time span may be annoying to the receiver. The subject may have then given much less attention to the feedback because they had already received 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 and performing well.

5.2.3 Exploratory Discussion

5.2.3.1 Noncompliance During Task Phases

Our main analysis is based on the assumption that the 333 respondents are all compliers, in which they all give an honest effort in answering the questions (no alternating or repeating answers) throughout all three task phases. For exploratory purposes, in this section, we consider a scenario in which we do include 46 noncompliers in our analysis. We will investigate whether there are signs of differential noncompliance.

We start with examining the distribution of noncompliers across control and treatment groups. We run a 5-sample proportions test to see whether the take-up rates across treatment groups are similar. With a p-value of 0.786036802694471, there are no statistically significant differences between take-up rates across groups at the p = 0.05 level and no evidence of differential noncompliance.

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

Table 13: 5-sample test for equality of proportions without continuity correction: `c(total_MedFeedback_compliers, total_selfreflect_compliers, total_positiveimages_compliers, out of c(total_MedFeedback_rows, total_selfreflect_rows, total_positiveimages_rows, total_negativeimages_compliers, total_control_compliers) out of total_negativeimages_rows, total_control_rows)`

Test statistic	df	P value	Alternative hypothesis	prop 1	prop 2	prop 3	prop 4	prop 5
1.726	4	0.786	two.sided	0.8684	0.88	0.8919	0.9091	0.8442

Now that we have calculated the take-up rates for each treatment group, 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 from our main regressions in Table 8. We see that the CACE estimates are slightly greater than the original regression estimates, with the exception of the Negative Images group, suggesting that the Negative Images were the least effective form of feedback in improving task performance.

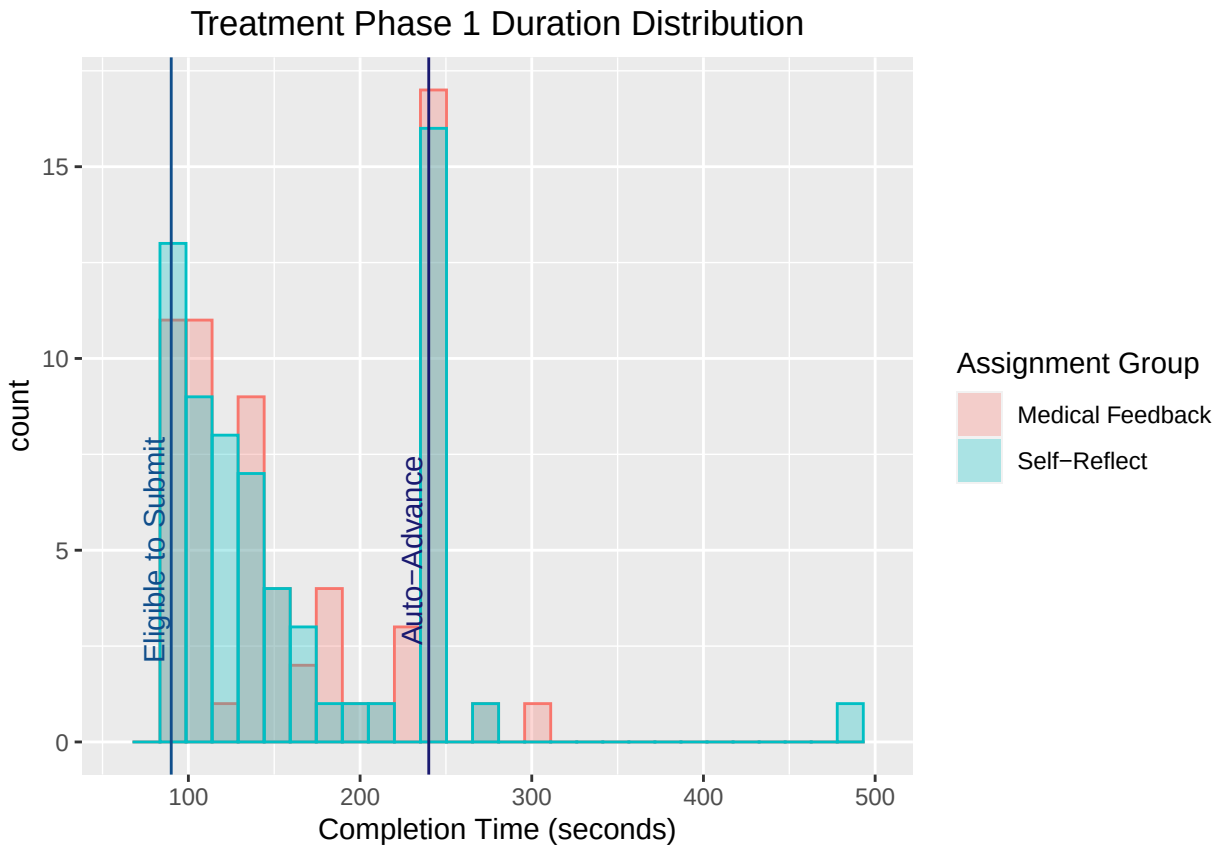
treatment_groups	list_CACEs	list_ITT	list_takeup
Any Treatment	0.0613	0.0544	0.8874
Medical Feedback	0.0766	0.0665	0.8684
Negative Images	0.0386	0.0351	0.9091
Positive Images	0.0647	0.0577	0.8919
Self-Reflect	0.0668	0.0588	0.88

5.2.3.2 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. However, 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. We find similar results when comparing the mean times of staying in treatment between Task Phase 1 and Task Phase 2. While not statistically significant, practically speaking, the results do show that for both Medical Feedback and Self-Reflect groups, the mean times of staying in treatment are shorter in Task Phase 2 than in Task Phase 1, which brings up the question of whether there were more noncompliers who breezed through the second round of treatment.


Treatment Phase 1 Duration Distribution
count
15
10
5
0
Eligible to Submit
Auto−Advance
100
200
300
400
500
Completion Time (seconds)
Assignment Group
Medical Feedback
Self−Reflect

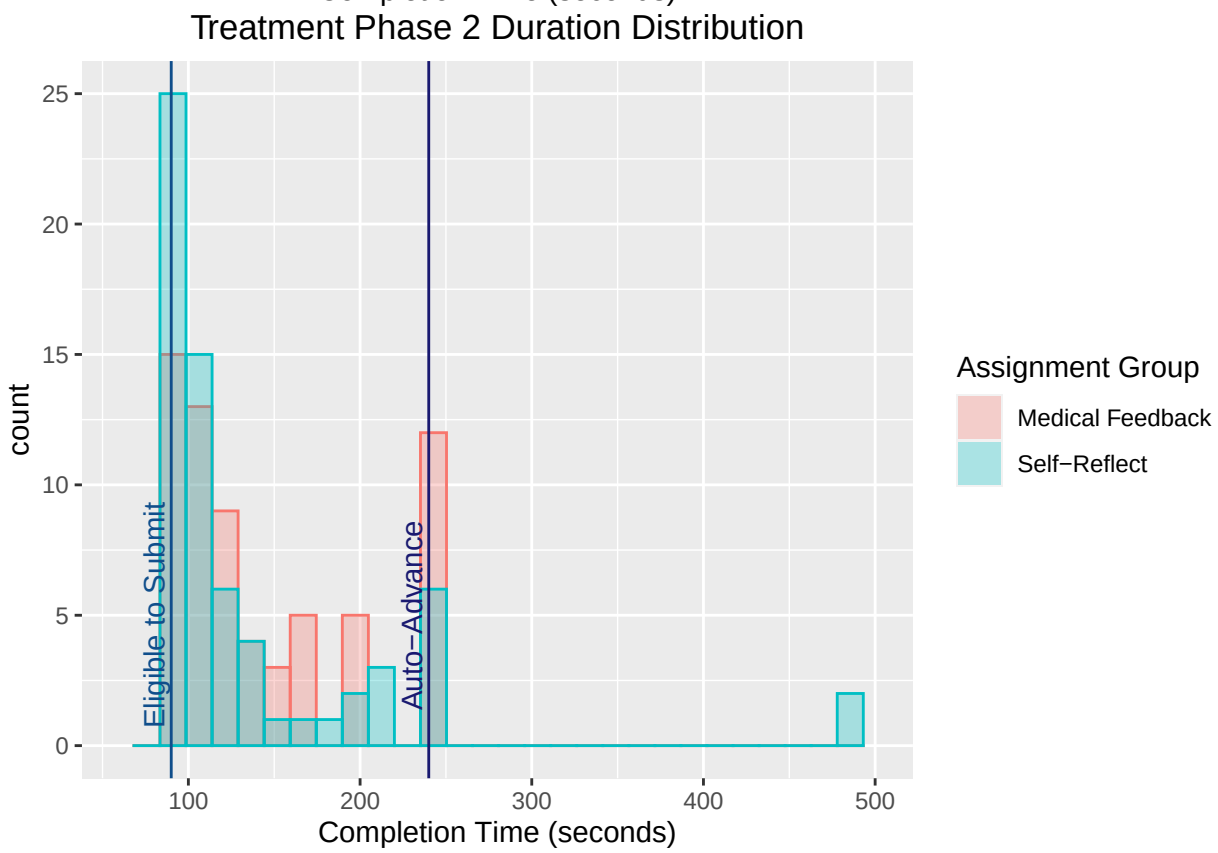

Treatment Phase 2 Duration Distribution
count
25
20
15
10
5
0
Eligible to Submit
Auto−Advance
100
200
300
400
500
Completion Time (seconds)
Assignment Group
Medical Feedback
Self−Reflect

Table 15: Welch Two Sample t-test: d_noncompliance_1[Assignment_Group %in% c("Medical Feedback") & and d_noncompliance_1[Assignment_Group %in% c("Self-Reflect") & Treatment_Phase1_SubmitTime <= Treatment_Phase1_SubmitTime <= 240, Treatment_Phase1_SubmitTime] and 240, Treatment_Phase1_SubmitTime]

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
1.545	84.66	0.1262	two.sided	134.5	123.1

Table 16: Welch Two Sample t-test: d_noncompliance_2[Assignment_Group %in% c("Medical Feedback") & and d_noncompliance_2[Assignment_Group %in% c("Self-Reflect") & Treatment_Phase2_SubmitTime <= Treatment_Phase2_SubmitTime <= 240, Treatment_Phase2_SubmitTime] and 240, Treatment_Phase2_SubmitTime]

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
0.8764	111	0.3827	two.sided	123.5	117.7

Table 17: Welch Two Sample t-test: d_noncompliance_1[Assignment_Group %in% c("Medical Feedback") & and d_noncompliance_2[Assignment_Group %in% c("Medical Feedback") & Treatment_Phase1_SubmitTime <= 240, Treatment_Phase1_SubmitTime] and Treatment_Phase2_SubmitTime <= 240, Treatment_Phase2_SubmitTime]

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
1.459	89.87	0.1482	two.sided	134.5	123.5

Table 18: Welch Two Sample t-test: d_noncompliance_1[Assignment_Group %in% c("Self-Reflect") & Treatment_Phase1_SubmitTime <= and d_noncompliance_2[Assignment_Group %in% c("Self-Reflect") & Treatment_Phase2_SubmitTime <= 240, Treatment_Phase1_SubmitTime] and 240, Treatment_Phase2_SubmitTime]

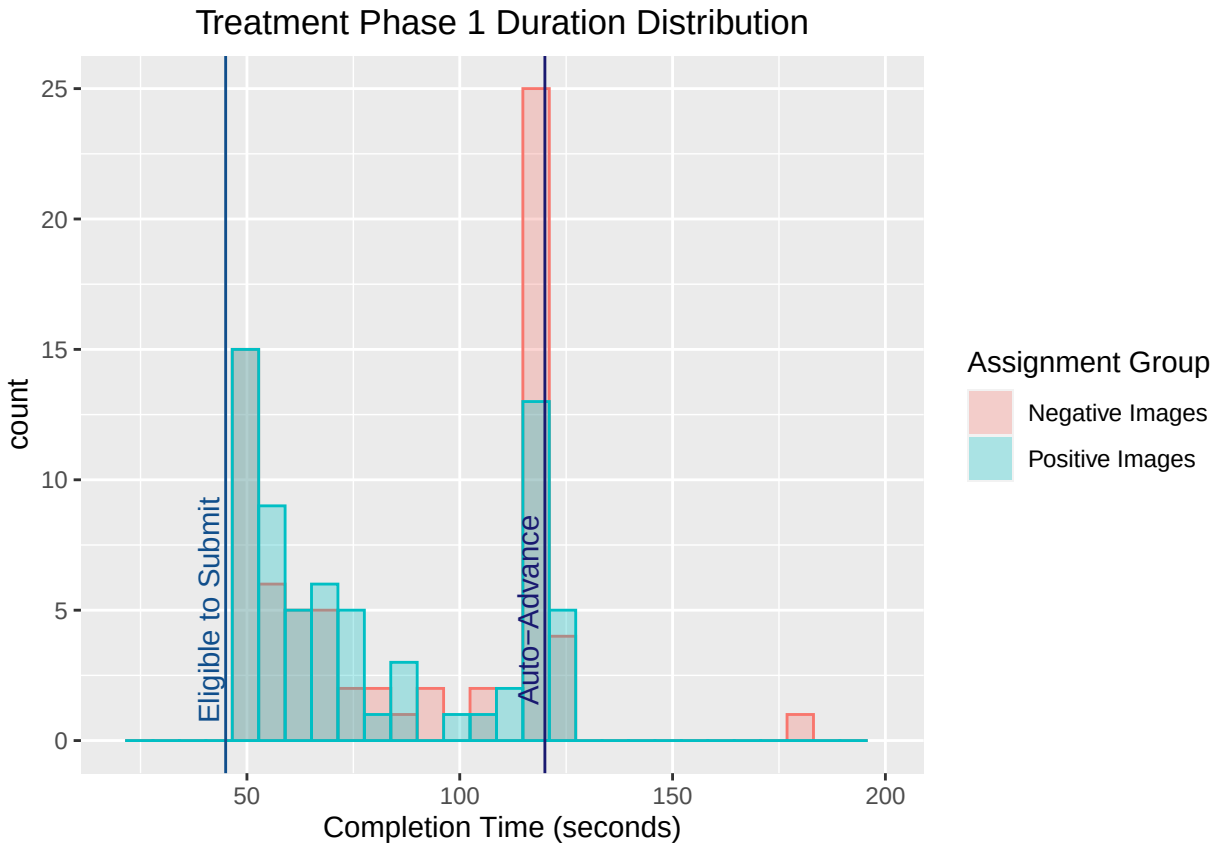
Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
0.8307	103.9	0.408	two.sided	123.1	117.7

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 during Treatment Phase 1. 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, where the difference is statistically significant at a p-value of 0.000158713618415726.

Similar to the Medical Feedback and Self-Reflect groups, in Treatment Phase 2, we see an 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, but the difference in mean times between Positive Images and Negative Images groups during this phase are not statistically significant at the p = 0.05 level.

While the differences in mean times are not statisically significant, we do see that the amount of time respondents in the Positive Images group spent in the second round of treatment was generally less than the first. However, respondents in the Negative Images group spent roughly equal amounts of time in the first and second rounds of treatment.



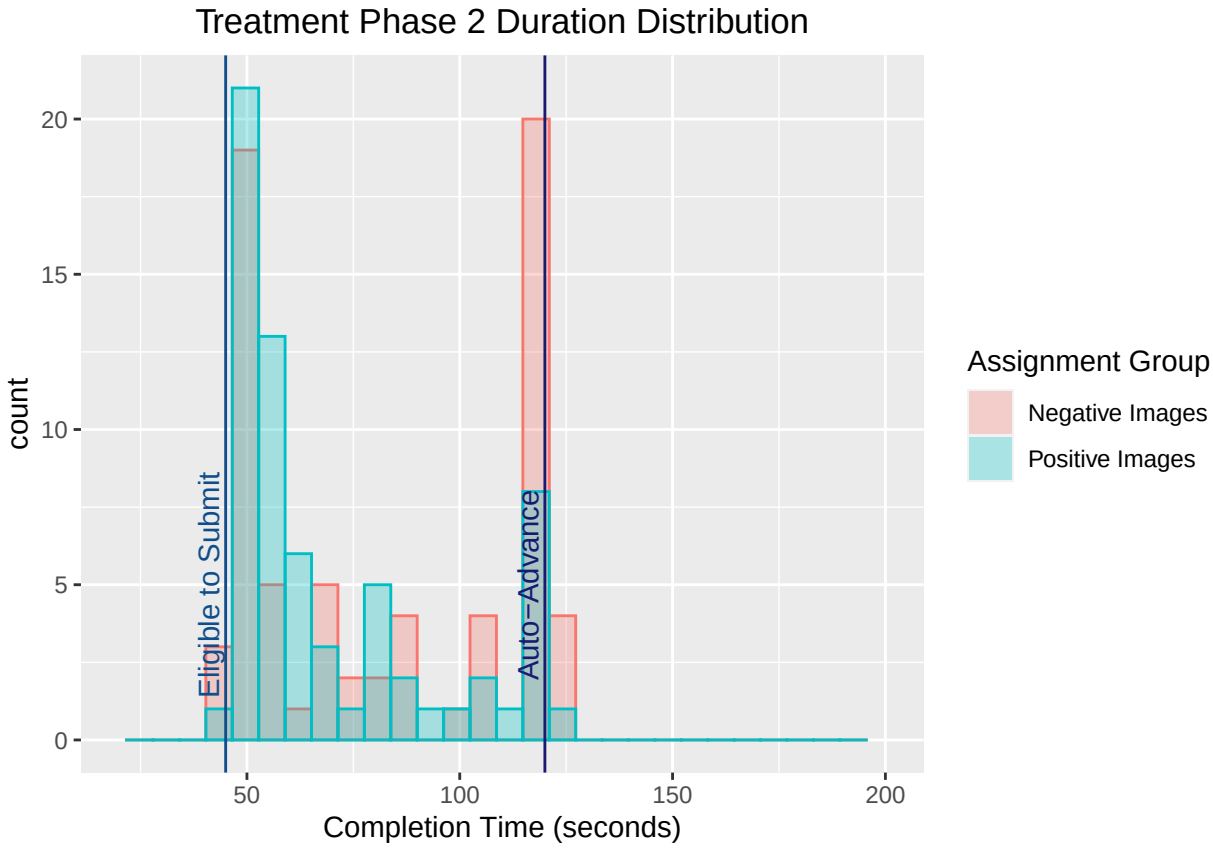


Table 19: Welch Two Sample t-test: `d_noncompliance_1[Assignment_Group %in% c("Positive Images")]` and `d_noncompliance_1[Assignment_Group %in% c("Negative Images") & Treatment_Phase1_SubmitTime] and Treatment_Phase1_SubmitTime <= 120, Treatment_Phase1_SubmitTime]`

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
3.921	103.5	0.0001587 * * *	two.sided	79.76	62.63

Table 20: Welch Two Sample t-test: `d_noncompliance_2[Assignment_Group %in% c("Positive Images")]` and `d_noncompliance_2[Assignment_Group %in% c("Negative Images") & Treatment_Phase2_SubmitTime] and Treatment_Phase2_SubmitTime <= 120, Treatment_Phase2_SubmitTime]`

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
1.384	109.1	0.1692	two.sided	69.7	63.79

Table 21: Welch Two Sample t-test: `d_noncompliance_1[Assignment_Group %in% c("Positive Images") &` and `d_noncompliance_2[Assignment_Group %in% c("Positive Images") & Treatment_Phase1_SubmitTime <= 120, Treatment_Phase1_SubmitTime]` and `Treatment_Phase2_SubmitTime <= 120, Treatment_Phase2_SubmitTime]`

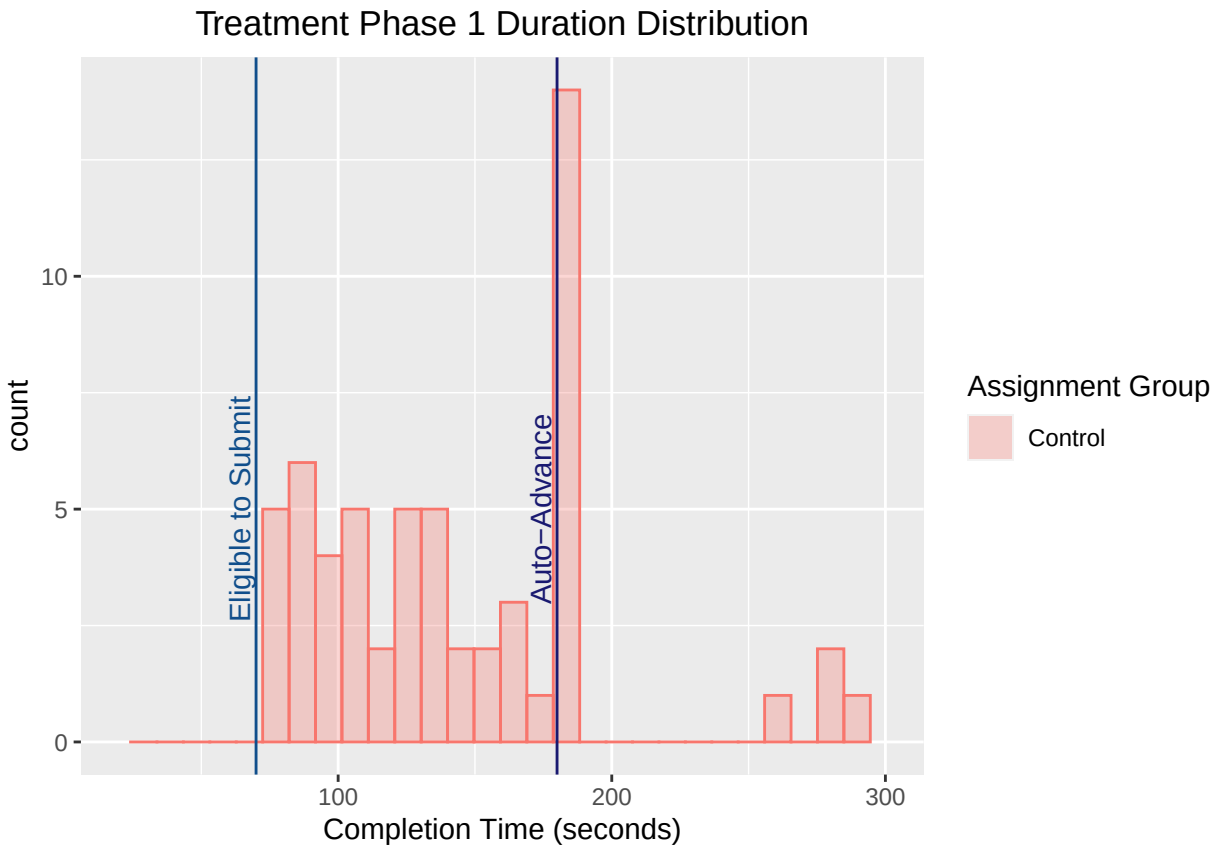
Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
1.401	97.39	0.1645	two.sided	66.71	61.71

Table 22: Welch Two Sample t-test: `d_noncompliance_1[Assignment_Group %in% c("Negative Images") &` and `d_noncompliance_2[Assignment_Group %in% c("Negative Images") & Treatment_Phase1_SubmitTime <= 120, Treatment_Phase1_SubmitTime]` and `Treatment_Phase2_SubmitTime <= 120, Treatment_Phase2_SubmitTime]`

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
-0.3008	83.78	0.7643	two.sided	62.63	63.79

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. However, this difference is not statistically significant with at the $p = 0.05$ level.


Treatment Phase 1 Duration Distribution
count
10
5
0
Eligible to Submit
Auto−Advance
100
200
300
Completion Time (seconds)
Assignment Group
Control

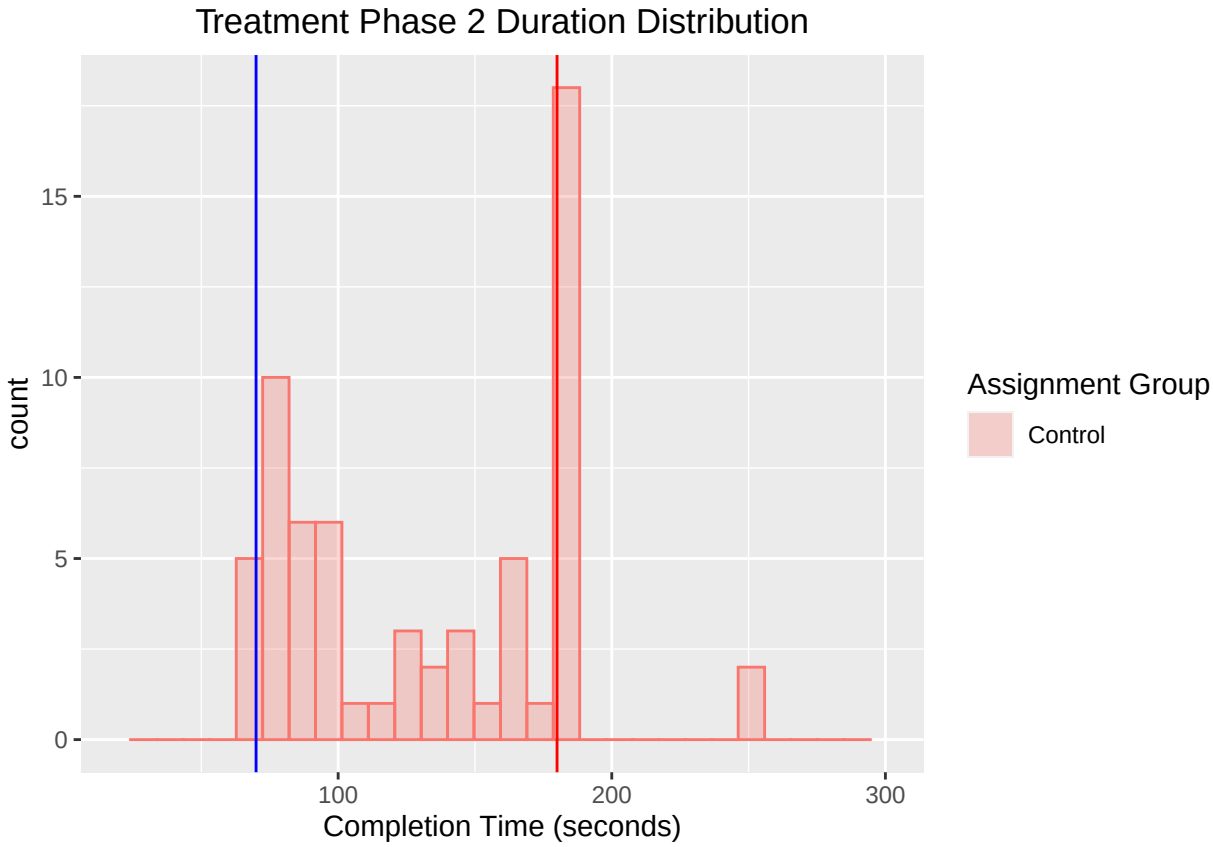

Treatment Phase 2 Duration Distribution
count
15
10
5
0
100
200
300
Completion Time (seconds)
Assignment Group
Control

Table 23: Welch Two Sample t-test: `d_noncompliance_1[Assignment_Group %in% c("Control") & Treatment_Phase1_SubmitTime <= 180, Treatment_Phase1_SubmitTime]` and `d_noncompliance_2[Assignment_Group %in% c("Control") & Treatment_Phase2_SubmitTime <= 180, Treatment_Phase2_SubmitTime]`

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
1.386	81.65	0.1697	two.sided	116.1	106.7

Throughout this exploratory discussion, with the exception of respondents in the Negative Images group, 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 what the treatment experience is like for respondents in the Negative Images group since respondents, on average, stayed in treatment relatively consistently. In regards to the other treatment groups in which the amount of time respondents are willing to spend in treatment decreases, it is plausible that this is a sign of feedback fatigue. Further research should be done on whether such behavior would trigger 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.

5.3 Power

```
 Two-sample t test power calculation

          n = 113
      delta = 0.06
         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 targeted medical feedback, 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.

6 Conclusions

Our experiment and following study shows that feedback contributes a statistically and practically significant effect in X-Ray analysis performance (ATE = 5.307, SE +- 2.352%). More specifically, targeted medical feedback saw the most statistically significant increases in performance (ATE = 5.882, SE +- 2.999), showing that expert opinion may lead to more significant outcomes in the real world. Along the same lines, self-reflection led to statistically and practically significant improvements on performance (ATE = 6.132, SE +- 3.068), 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 = 4.355, p = 0.118), 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 = 0.366, p = 0.857) 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

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).

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.

7 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.

To combat noncompliance issues, we can redesign the Qualtrics survey flow so that we can stop the survey for potential noncompliers when they either give alternating answers or the same answers during Task Phase 1. This will help ensure that only complying respondents get randomized into treatment. As for dealing with potential noncompliance during the Treatment Phases, other design considerations would need to be included so that respondents are encouraged to stay within the treatment phase for longer time periods since we noticed that nearly all respondents in treatment groups spent less time in Treatment Phase 2 than in Treatment Phase 1. Other 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 (in case respondents

rush through answering questions). We saw several terse responses to open-ended questions in the self-reflect and control groups, which are possible signs of noncompliance. Implementing a word minimum limit for respondents in the survey would deter this behavior. Furthermore, we noticed that there were respondents who were able to stay in treatment for longer times than originally prescribed. This could be due to a technical issue from Qualtrics that would have to undergo further testing.

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 that were meant to parallel the real-world. 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 such nuances of providing feedback, and potentially help us determine if the presence of a writing component, a reading component, a visual component, an answer key, or an interaction of multiple factors would best improve task performance.