Mid-Test Effects

Contents

1 Introduction

This document estimates the effect of immediate versus delayed feedback on the mid-test (total math score) in the FH2T RCT. It is based on the 1589 subjects initially randomized between the two conditions, 795 randomized to the Instant condition, and 794 randomized to the delayed feedback condition. Randomization was blocked within the 169 classrooms.

The following section discusses attrition–students who did not take the mid-assessment.

The next section estimates treatment effects.

2 Attrition

When randomized subjects do not have outcome information in an RCT, effect estimates may be biased. Randomization creates treatment groups that, in expectation, are identical in measured and unmeasured characteristics, referred to as "covariate balance." However, attrition is not controlled by the researcher, nor is it random; in particular, if different types of subjects attrit in different treatment groups, the remaining subjects will not necessarily be comparable across groups. That said, if the overall level of attrition is low, or if the level of attrition is similar between the two treatment groups, and if important covariates remain balanced between treatment groups after excluding attritors, the bias might be negligible.

The What Works Clearinghouse publishes standards for attrition bias based on the level of overall attrition as well as differential attrition–the difference in attrition levels between treatment groups.

In practice, we will omit the 710 students without mid-test scores (and in some analyses, also the additional 26 students without pre-test scores). In 43 classrooms, no students took the mid-test, and in 2 additional classrooms, no student took both the pre- and the mid-test. In 15 classrooms, there were no students with pre-test scores in one of the two treatment groups; these classrooms were dropped from all analyses other than Model 1, below. Dropping these entire classrooms enhances the validity of the analyses, but reduces the sample size by 23 students.

2.1 Attrition Rates

The overall attrition for the mid-test was 45%–710 students out of 1589 did not complete the mid-test.

Table 1: Attrition by treatment group.

	Delay	Instant
Attrit	45.3	44
Took Mid-Test	54.7	56

Table 1 gives attrition by treatment group. The differential attrition was -1.31%. Taken together, the (high) overall attrition with the (low) differential attrition means that this study meets the conservative WWC standards.

An important subgroup is composed of the 1223=77% of students who have pre-test scores. Among this subgroup, 853 also took the mid-test, so attrition is much lower: 30%. The differential attrition is also acceptable: -0.45%.

2.2 Attrition and Covariates

2.2.1 Who Attritted?

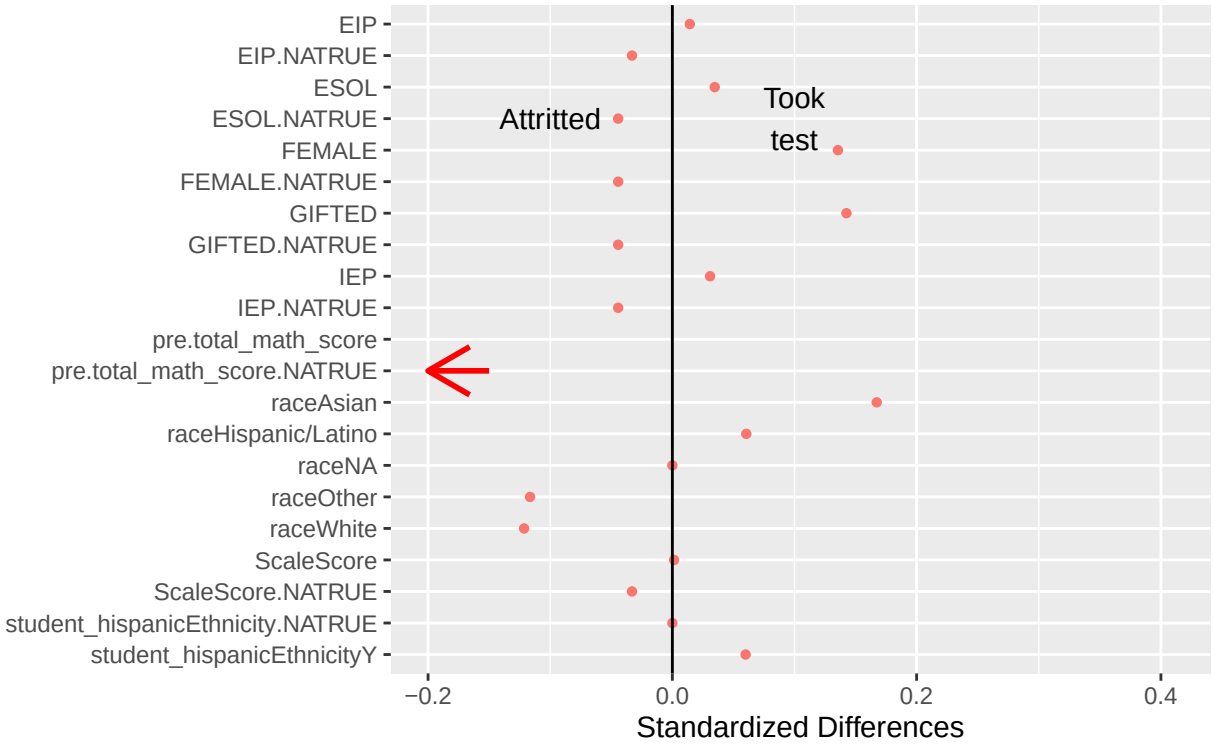

Figure 1: Standardized differences of covariate means between students who took the mid-test and those who didn't

Who attritted? Figure 1 shows standardized differences of covariates between students who attritted and those who took the mid-test. Students who took the test had higher pretest and 5th-grade scale scores, were more likely to be gifted, and more likely to be Asian that students who attritted. Students who attritted were much less likely to have taken the pre-test.

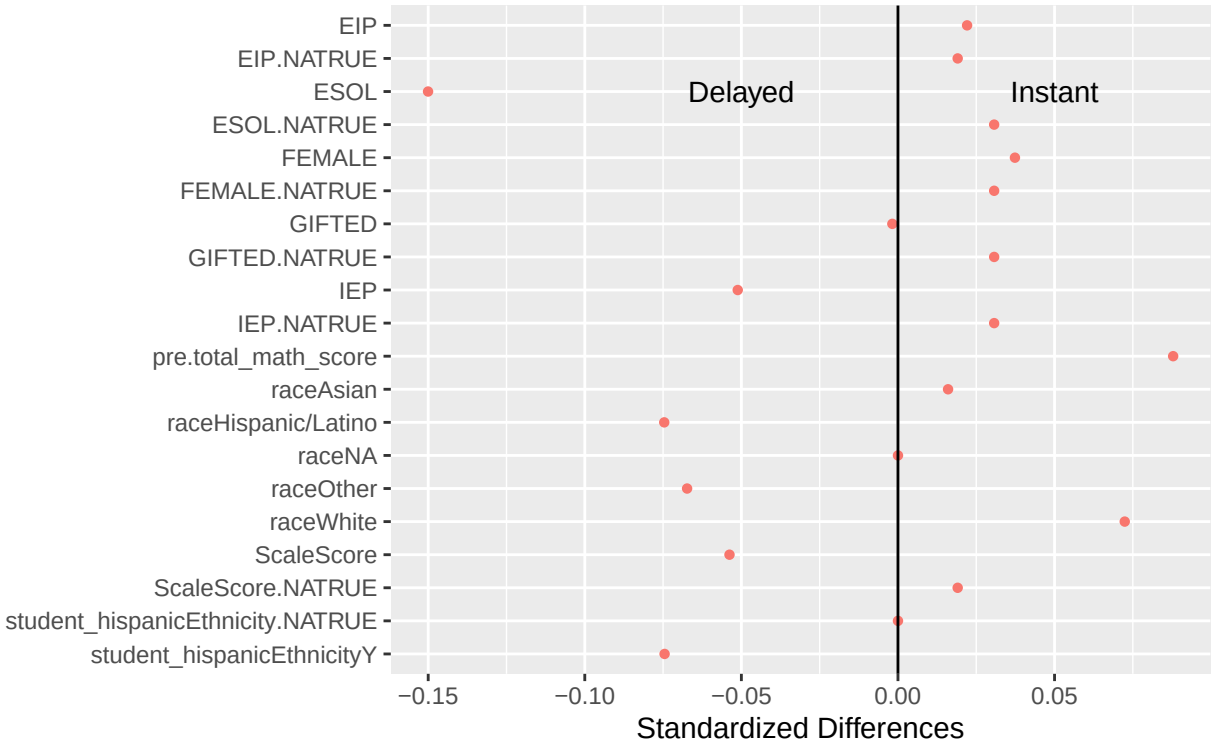

Figure 2: Standardized differences of covariate means between Instant and Delayed-feedback students among non-attritors

2.2.2 Covariate Balance among Test-Takers

Are instant- and delayed-feedback non-attritors comparable? Figure 2 compares the two treatment groups among those who took the mid-test. A p-value testing overall balance was $p=$0.205, meaning that the groups were more balanced than 21% of randomized experiments.

The only notable difference was that delayed feedback students who took the mid-test tend to be more likely to be ESOL than those in the instant-feedback condition.

These results suggest that confounding bias due to attrition is unlikely to be a major concern. Nevertheless, we adjust for pre-treatment covariates in one of the models below.

3 Growth

```
##
##  One Sample t-test
##
## data:  smallDat$mid.total_math_score - smallDat$pre.total_math_score
## t = 8.4563, df = 831, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.4863890 0.7804379
## sample estimates:
## mean of x
## 0.6334135
```

4 Treatment Effects

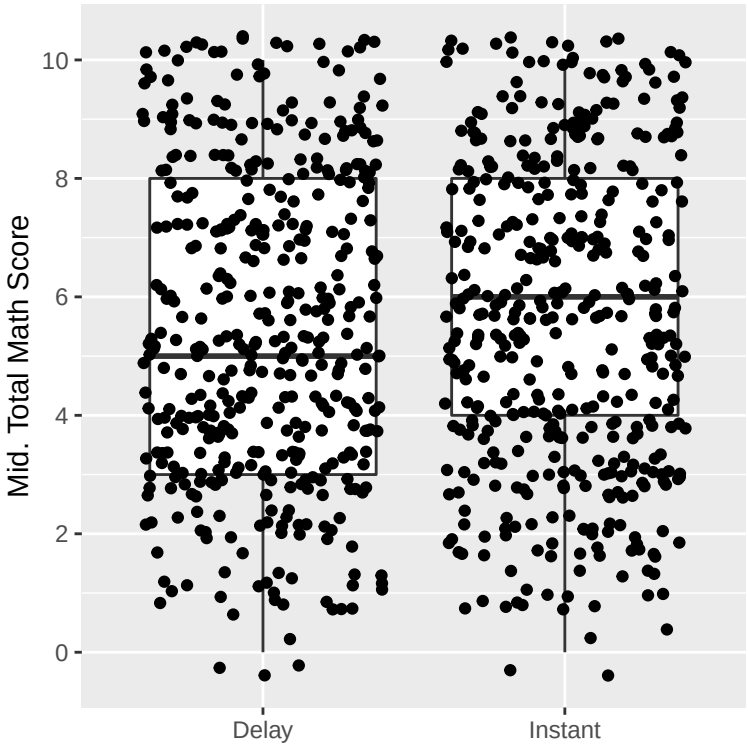

Figure 3: Boxplots of Mid-test scores for the two treatment conditions, with jittered scores

Figure 3 gives boxplots of the total scores for the two treatment conditions, with individual scores plotted as jittered points. The median score for Instant-feedback students is one point higher than the median for delayed-feedback students, but there is wide variation in both groups.

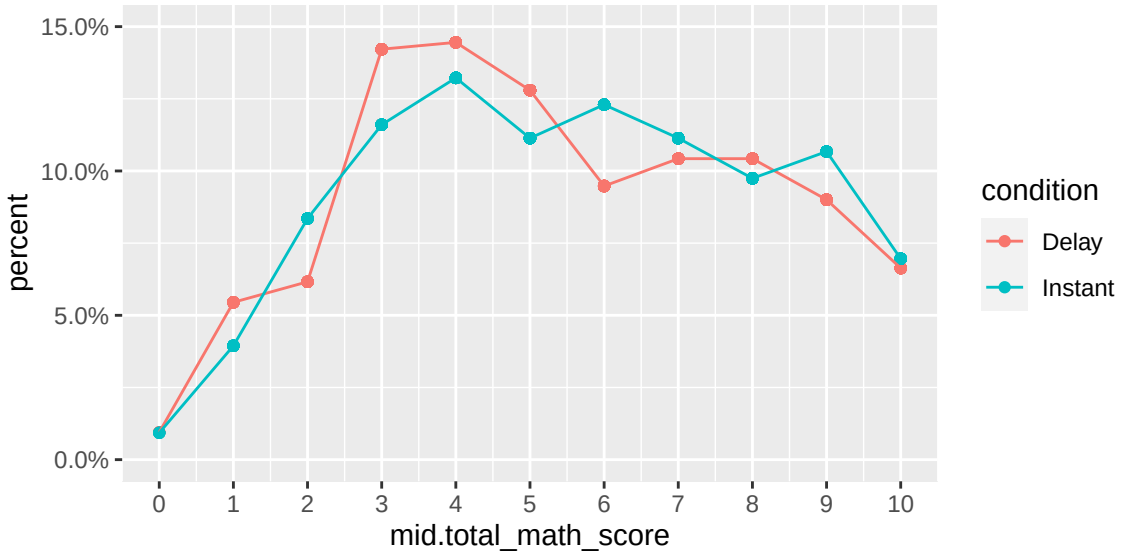

Figure 4: The percent of each treatment group achieving each score on the mid-test

Figure 4 shows the percent of students in each treatment group who achieved each possible score on the mid-test. A slightly higher percentage of delayed-feedback students scored 3, 4, or 5, and a slightly higher percentage of instant-feedback students scored 6 or 9.

Figure 5 gives a scatterplot of pre-test vs post-test scores with separate OLS fits in each treatment group. The best-fit lines are almost identical, but not quite–ignoring statistical error, it appears that students with very high pre-test scores did better on the mid-test if they were in the Instant condition, while students with very low pre-test scores did better on the mid-test if they were in the Delay condition. We can test this with an OLS model below.

Table 2 give the results of OLS models estimating the effect of assignment to the instant- versus delayed-feedback condition on mid-tests. Model 1 is just a comparison of the mean test scores. Models 2-4 also include fixed-effects for classroom (the randomization blocks). These estimate a weighted-average treatment effect, where the weights are chosen to maximize precision. Models 3-4 adjust for covariates, and include only students with pretest scores. All standard errors are heteroskedasticity-consistent, estimated using the `estimatr` package in `R`.

```
## [[1]]
##                          2.5 %    97.5 %
## conditionInstant -0.01910053 0.5007302
##
## [[2]]
##                         2.5 %   97.5 %
## conditionInstant -0.1021504 0.404992
##
## [[3]]
##                          2.5 %    97.5 %
## conditionInstant -0.06477161 0.4348674
```

```
## [[1]]
##                                  2.5 %    97.5 %
```

	Model 1	Model 2	Model 3	Model 4	Model 5
conditionInstant	0.24$^{\cdot}$	0.15	0.19	0.16	0.19
	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)
pretest		0.36***	0.30***	0.33***	0.27***
		(0.04)	(0.05)	(0.05)	(0.05)
ESOL1		−0.04	0.46	−0.07	0.44
		(0.32)	(0.36)	(0.32)	(0.36)
ScaleScore			0.43***		0.42***
			(0.11)		(0.11)
EIP1			−0.25		−0.26
			(0.31)		(0.31)
IEP1			−0.16		−0.17
			(0.24)		(0.24)
FEMALE1			0.16		0.16
			(0.14)		(0.14)
GIFTED1			0.37$^{\cdot}$		0.37$^{\cdot}$
			(0.20)		(0.20)
raceHispanic/Latino			−0.23		−0.23
			(0.25)		(0.25)
raceAsian			0.47$^{\cdot}$		0.48$^{\cdot}$
			(0.26)		(0.26)
raceOther			0.08		0.08
			(0.26)		(0.26)
ScaleScoreMISSTRUE			0.16		0.16
			(1.04)		(1.02)
ESOLMISSTRUE			−0.15		−0.15
			(1.07)		(1.05)
raceMISSTRUE			0.36		0.38
			(0.53)		(0.53)
conditionInstant:pretest				0.06	0.06
				(0.05)	(0.05)
R^2	0.55	0.59	0.61	0.59	0.61
Adj. R^2	0.47	0.52	0.54	0.52	0.54
Num. obs.	853	853	853	853	853
RMSE	1.89	1.80	1.76	1.80	1.76

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Table 2: OLS models estimating the effect of assignment to the instant feedback condition versus the delayed feedback condition. All models except 11 include fixed-effects for classroom; models 3 & 4 only include students with pretest scores. Confidence intervals are in brackets under coefficient estimates.

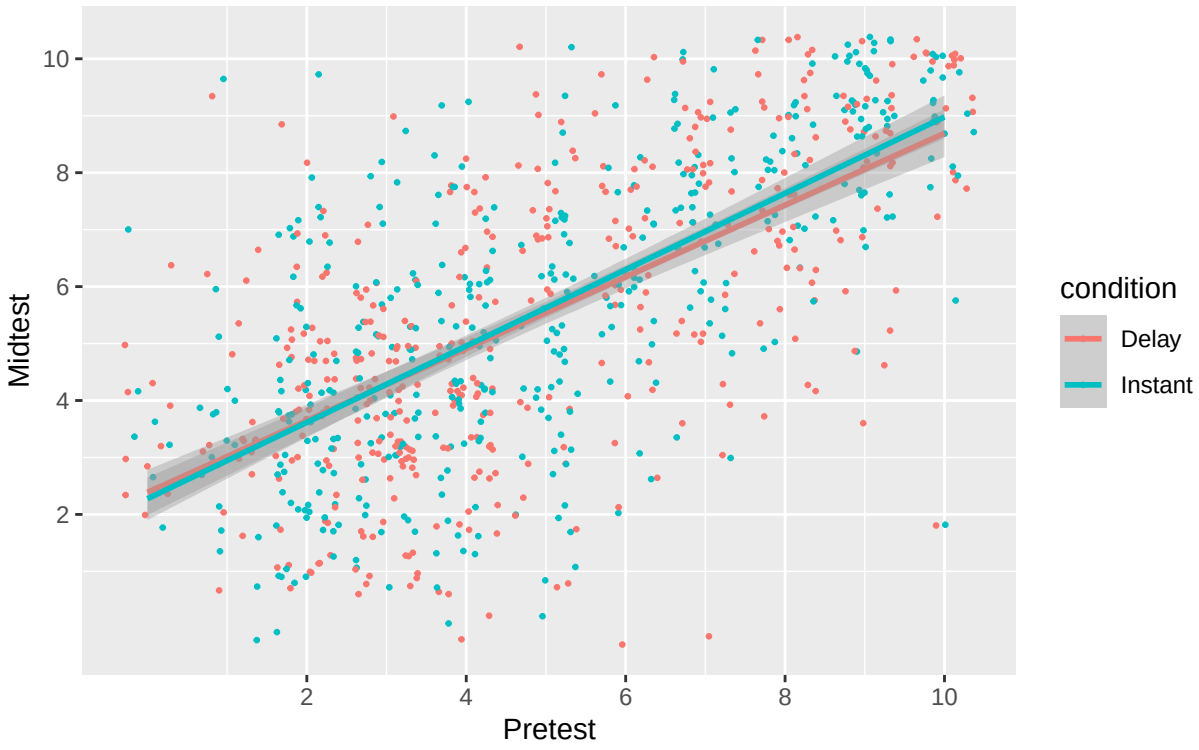

Figure 5: Scatterplot of pre- and mid-test scores, with jitter to avoid overplotting and separate OLS fits by treatment group.

```
## conditionInstant          -0.09622617 0.4102985
## conditionInstant:pretest -0.03390006 0.1458155
##
## [[2]]
##                                2.5 %    97.5 %
## conditionInstant          -0.05855839 0.4400994
## conditionInstant:pretest -0.02913099 0.1486109
```

All of the models estimated a positive effect for instant- versus delayed-feedback, but none of them estimated significant effects–all models agree that the instant versus delayed feedback affects average scores by at most half of a point.

Model 7 estimated a positive interaction between the effect of being assigned to the Instant condition and pretest scores–i.e. that the effect becomes larger (more positive) as pretest scores increase–however, the data are also consistent with a 0 or slightly negative interaction (p=0.187). The model predicts an effect of 0.012, (CI: [-0.366,0.389]) for subjects with pretest scores of 2, roughly 1 standard deviation below the mean and 0.31, (CI: [0.014,0.607]) for subjects with pretest scores of 7, roughly 1 standard deviation above the mean.