What is an RCT experiment?

A Randomized Controlled Trial (RCT) is a type of scientific experiment, often used in fields like medicine, social sciences, and policy evaluation, to determine the effectiveness of an intervention or treatment. In this specific I am using it for the purpose of policy evaluation, whereby an RCT aims to rigorously assess the impact of a new policy or program on a specific outcome.

There are several core components of an RCT. Namely, the participants that are the focus of the experiments. The policy that is being tested. Control and Treatment groups, and the outcome measure which are the specific indicators that will be used to assess and quantify the impact of the RCT experiments.

Why is an RCT experiment necessary for policy evaluation

The primary goal of randomization is to create groups that are statistically similar in all relevant characteristics at the start of the experiment. By randomly assigning participants, the aim is to distribute these characteristics evenly across the treatment and control groups.

One key aspect of an RCT is in its ability to isolate the effects of the treatment, which is done by comparing the outcomes of the treatment group with the outcomes of the control group. If the treatment group shows a statistically significant improvement in the outcome compared to the control group, we can be more confident that this improvement is attributable to the intervention. The random nature of the experiment also reduces bias and ensures the validity of the results.

A well-conducted RCT with a randomized control group provides the strongest evidence for establishing a causal relationship. By ensuring the groups are similar at the start and only differ in whether they received the intervention, any significant difference in outcomes can be attributed to the policy with a higher degree of confidence.

Argument 1: We can evaluate our program by comparing student scores of students who receive the extra tuition with students who didn't receive the tuition.

The primary flaw in this argument is the fact that the two groups are not equivalent at the outset. The students who receive extra tuition are specifically chosen because they scored lower on their mathematics exam. This creates an inherent difference between the treatment group and the comparison group. The point of this experiment is to isolate the impact of the policy regardless of the characteristics of the groups. The lower-scoring

students might have different levels of prior knowledge, motivation, learning styles, parental support, or face other challenges that contribute to their lower scores. Simply comparing their scores after the tuition program concludes will likely reflect these pre-existing differences as well as any potential effect of the tuition, which would make it impossible to isolate whether any observed difference in scores is due to the extra tuition or some other hidden disparities.

For example, even if the treated group shows improvement, it might still score lower than the untreated group, and this difference could be wrongly attributed to the ineffectiveness of the tuition rather than their initial lower baseline. Conversely, if the treated group shows a greater improvement, it might be due to them having more room for improvement initially.

Additionally, If the group selected based on the last exam is indeed more variable, randomization becomes even more crucial. It helps to ensure that this variability is evenly distributed between the treatment and control groups. This allows us to see if the tuition has a general effect on students who recently underperformed, regardless of the underlying reasons for that underperformance.

Argument 2: We can evaluate our program simply by comparing the before and after scores of the students who received the program with those students who didn't receive the program.

Even with before-and-after comparisons for both groups, if the initial groups are not comparable, then any differences in the *change* in scores might still be driven by those initial differences rather than the tuition itself. For example, the lower-performing group might show a larger increase simply because they started from a lower point.

When students are selected based on a single low score on their last exam, we are likely to include students who might have experienced a temporary dip in their performance. Statistically, these students are likely to score higher on subsequent exams simply due to natural variation and not necessarily because of the extra tuition. This phenomenon, known as regression to the mean, could lead us to falsely attribute any improvement in their scores to the extra tuition when it might have happened anyway. The control group, not selected based on this recent low score, wouldn't be subject to the same degree of regression to the mean, making a direct comparison of score changes misleading.

In this scenario, pre-treatment balance refers to the similarity of the students in the different treatment groups and the control group before they are exposed to the treatment. Ideally,

the outcome of this task should show how statistically similar both the groups are in terms of their baseline characteristics, in this case their mathematics scores.

Pre-treatment balance in our main outcome of interest, mathematics scores, is crucial because it allows to confidently attribute any significant differences observed in post-treatment scores between the groups to the extra tuition programs (or the lack thereof in the control group). If the groups are not balanced at the baseline in their mathematics abilities, any improvements or differences we see after the tuition could be due to these pre-existing disparities rather than the effect of the intervention itself. A good pre-treatment balance ensures that the groups start on a level playing field with respect to their mathematical skills, making it possible to isolate the causal impact of the different tuition approaches on their subsequent performance.

In order to get a cursory understanding of where the current pre-treatment balance lies for this experiment between the treatment and control groups, descriptive statistics and visual inspection is going to be used.

Descriptive Statistics

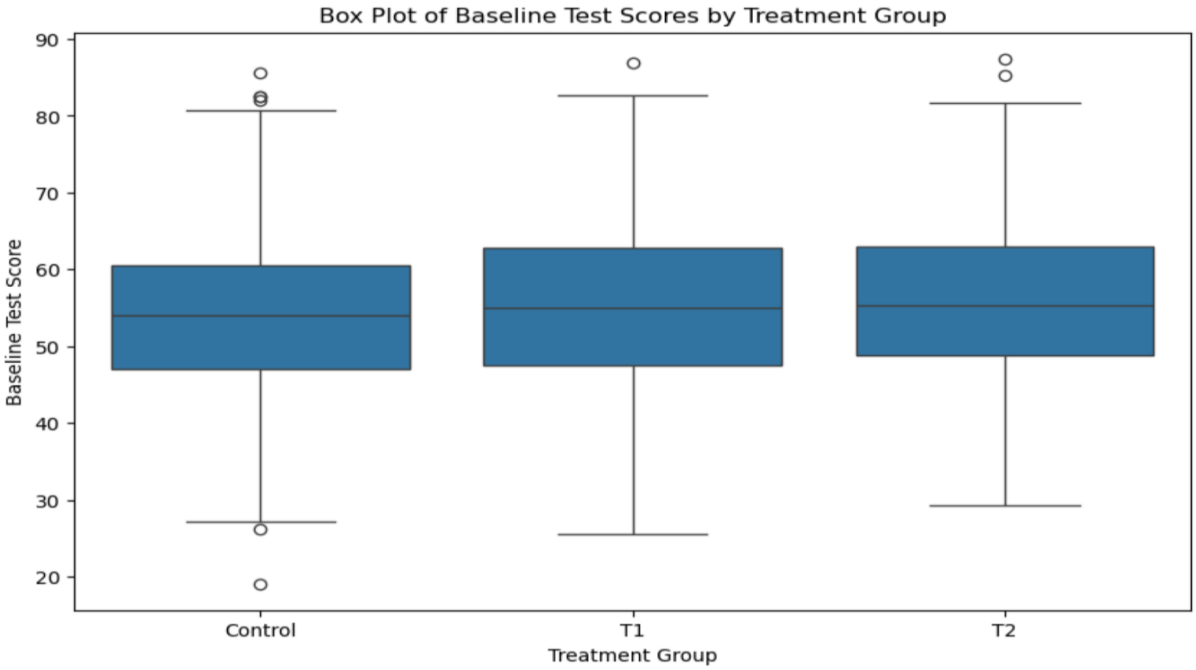
Treatment	count	mean	std	min	25%	50%	75%	max
Control	400	53.97	10.36	19.08	46.98	54.02	60.54	85.57
T1	400	55.09	10.89	25.56	47.46	55.06	62.77	86.87
T2	400	55.93	9.93	29.28	48.8	55.29	62.92	87.44

They are quite close to each other. While there's a slight increase from the Control group to T1 and then to T2, the differences are relatively small. This suggests that, on average, the groups started at a similar level of mathematics proficiency.

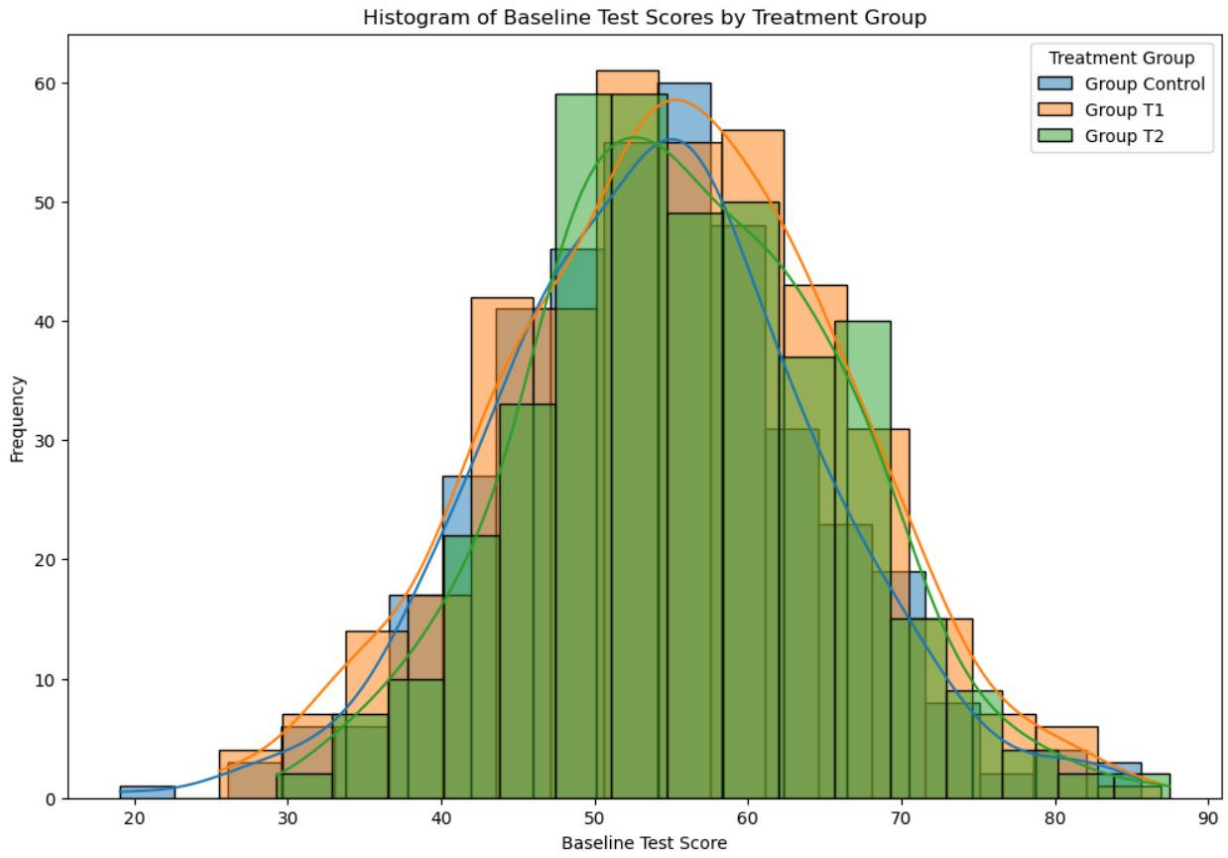
The standard deviations are also quite similar, indicating that the spread or variability of baseline scores within each group is comparable. This suggests that the level of dispersion around the average score is not drastically different across the groups.

The minimum scores range from approximately 19 to 29, and the maximum scores range from approximately 85 to 87 across the groups. This suggests a similar overall range of student abilities in mathematics at baseline in all three groups.

Visual Inspection



Visually, the box plot supports the findings from the descriptive statistics. The distributions of baseline test scores for the Control group, T1, and T2 appear to be quite similar in terms of their central tendency, spread, and the presence of outliers. The second treatment group does seem to be fair marginally well compared to the control group, which also has a few outliers on the lower side, while neither of the treatment groups show any outliers on that side. Overall, however, this provides confidence that the randomization process was effective in creating comparable groups before the extra tuition program can be implemented.



The histogram shows a high degree of overlap in the distributions, similar central tendencies, and comparable spread of the baseline test scores across the groups. This along with the previous results provides strong evidence of a good pre-treatment balance. This suggests that the randomization process was successful in creating groups that were statistically similar in terms of their baseline mathematics performance before the intervention began. The first treatment group does seem like it has a slightly wider distribution than the others. Therefore, we can proceed with the analysis of the post-treatment outcomes to evaluate the effectiveness of the extra tuition programs with a reasonable degree of confidence that any observed differences are likely attributable to the interventions rather than pre-existing disparities between the groups.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     test_score_endline   R-squared:                       0.796
Model:                            OLS   Adj. R-squared:                  0.795
Method:                 Least Squares   F-statistic:                     1552.
Date:                Wed, 19 Mar 2025   Prob (F-statistic):               0.00
Time:                        01:41:00   Log-Likelihood:                -3846.1
No. Observations:                1200   AIC:                             7700.
Df Residuals:                    1196   BIC:                             7721.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      8.6600      0.945      9.163      0.000       6.806      10.514
T1             4.7431      0.423     11.213      0.000       3.913       5.573
T2            -8.7270      0.424    -20.590      0.000      -9.559      -7.895
test_socre_baseline     1.0191      0.017     61.350      0.000       0.987       1.052
==============================================================================
```

The estimated ANCOVA model explains approximately 79.6% of the variance in endline test scores (R-squared = 0.796), indicating a good fit to the data.

Interestingly all the results are statistically significant with a p value of 0.000.

Intercept (8.66): This is the estimated endline test score for a student in the Control group with a baseline test score of zero, assuming all other predictors are also at their reference level. Which implies that some performance increase is expected.

Treatment 1 (Coefficient = 4.74): The coefficient for Treatment 1 is positive and statistically significant. This suggests that, on average, students in Treatment group 1 scored 4.74 points higher on the endline test compared to students in the Control group, after accounting for differences in their baseline test scores. This is in line with our general assumptions about the impact of the treatment. Given higher tuition scores will increase.

Treatment 2 (Coefficient = -8.73): The highly significant negative coefficient for Treatment 2 is a noteworthy and unexpected finding. Students in Treatment group 2 performed substantially worse on the endline test compared to the Control group, even after controlling their baseline scores. This result warrants further investigation. It could indicate that the specific intervention in Treatment 2 was not effective, or perhaps even had a detrimental effect on student performance. It's important to consider the nature of the intervention in Treatment 2 to understand potential reasons for this negative correlation.

Baseline Test Score (Coefficient = 1.02): This is also positive and also highly statistically significant with $p < 0.001$. This indicates a strong positive relationship between baseline and endline test scores. For every one-point increase in the baseline test score, the endline test score is estimated to increase by 1.02 points, after accounting for the treatment groups.

OLS Regression Results

Dep. Variable:	test_score	R-squared:	0.205
Model:	OLS	Adj. R-squared:	0.201
Method:	Least Squares	F-statistic:	47.46
Date:	Wed, 19 Mar 2025	Prob (F-statistic):	3.50e-109
Time:	14:43:54	Log-Likelihood:	-9211.7
No. Observations:	2400	AIC:	1.845e+04
Df Residuals:	2386	BIC:	1.853e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	53.3272	0.864	61.729	0.000	51.633	55.021
C(school_id)[T.2.0]	1.5457	0.920	1.680	0.093	-0.259	3.350
C(school_id)[T.3.0]	-0.0357	0.921	-0.039	0.969	-1.841	1.770
C(school_id)[T.4.0]	2.5984	0.921	2.820	0.005	0.792	4.405
C(school_id)[T.5.0]	2.5730	0.921	2.794	0.005	0.767	4.379
C(school_id)[T.6.0]	-0.3343	0.921	-0.363	0.717	-2.140	1.471
C(school_id)[T.7.0]	1.5941	0.920	1.732	0.083	-0.211	3.399
C(school_id)[T.8.0]	-0.0781	0.920	-0.085	0.932	-1.883	1.727
C(class_id)[T.2.0]	-0.7033	0.460	-1.529	0.127	-1.606	0.199
T1	1.1303	0.797	1.418	0.156	-0.433	2.694
T2	1.9862	0.799	2.487	0.013	0.420	3.552
post_binary	9.6927	0.797	12.162	0.000	8.130	11.255
post_binary:T1	4.7646	1.127	4.227	0.000	2.554	6.975
post_binary:T2	-8.6897	1.127	-7.710	0.000	-10.900	-6.480

The statistically significant coefficient for post binary indicates that, on average, the test scores for the control group increased by approximately 9.69 points from the pre-treatment period to the post-treatment period, after accounting for school and class fixed effects. This suggests a generally positive time trend in test scores for the control group.

The coefficients for the treatment dummies in the pre-treatment period are:

T1: 1.1303 (p=0.156), which is not statistically significant at the 0.05 level.

T2: 1.9862 (p=0.013), which is statistically significant. This indicates that Treatment Group 2 had a significantly higher baseline test score (by about 1.99 points) compared to the control group.

The important features of the results are, however, given by the interaction terms:

post_binary:T1: 4.7646 (p<0.001). This coefficient represents the additional impact of Treatment 1 on the test scores in the post-treatment period compared to the control group's change over time. It suggests that Treatment Group 1 experienced an additional increase of about 4.76 points in their test scores due to the treatment.

post_binary:T2: -8.6897 (p<0.001). This coefficient represents the additional impact of Treatment 2 on the test scores in the post-treatment period compared to the control group's change over time. It suggests that Treatment Group 2 experienced a decrease of about 8.69 points in their test scores due to the treatment, relative to the control group's improvement, this is in line with the results from the ANCOVA results from before.

The coefficients for the school fixed effects indicate that there are statistically significant differences in average test scores across some schools compared to the reference school. The coefficient for the class fixed effect is not statistically significant at the 0.05 level.

OLS Regression Results

Dep. Variable:	test_score_endline	R-squared:	0.851
Model:	OLS	Adj. R-squared:	0.850
Method:	Least Squares	F-statistic:	618.9
Date:	Wed, 19 Mar 2025	Prob (F-statistic):	0.00
Time:	11:14:53	Log-Likelihood:	-3657.7
No. Observations:	1200	AIC:	7339.
Df Residuals:	1188	BIC:	7400.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.8197	0.911	0.899	0.369	-0.968	2.608
C(school_id)[T.2.0]	0.2734	0.592	0.462	0.644	-0.888	1.435
C(school_id)[T.3.0]	0.3850	0.592	0.651	0.515	-0.776	1.546
C(school_id)[T.4.0]	0.2279	0.592	0.385	0.700	-0.934	1.390
C(school_id)[T.5.0]	1.3615	0.593	2.297	0.022	0.199	2.524
C(school_id)[T.6.0]	0.0299	0.592	0.051	0.960	-1.131	1.191
C(school_id)[T.7.0]	-0.0172	0.592	-0.029	0.977	-1.179	1.144
C(school_id)[T.8.0]	0.2142	0.592	0.362	0.717	-0.947	1.375
C(class_id)[T.2.0]	0.0379	0.296	0.128	0.898	-0.543	0.619
T1	12.0396	0.374	32.198	0.000	11.306	12.773
T2	13.6138	0.354	38.457	0.000	12.919	14.308
test_score_baseline	1.0028	0.014	70.232	0.000	0.975	1.031

The ANCOVA model shows a very high R-squared of 0.850, indicating that 85% of the variance in endline test scores is explained by the treatment groups and the baseline test score.

The coefficient for test_score_baseline is 1.0051, which is highly statistically significant at p <0.001. Suggesting a strong positive relationship between the baseline and endline test scores.

The coefficients for the treatment groups, relative to the control group, are as follows:

T1 is 12.0406 (p<0.001). This indicates that, on average, students in Treatment Group 1 scored about 12.04 points higher on the endline test compared to the control group, after controlling their baseline test scores.

T2 is 13.6146 (p<0.001). This indicates that, on average, students in Treatment Group 2 scored about 13.61 points higher on the endline test compared to the control group, after controlling for their baseline test scores. This is a turnaround from the non-stratified dataset, which had a negative coefficient.

Both Treatment Group 1 and Treatment Group 2 had a statistically significant positive impact on students' endline test scores. This indicates that both interventions were effective in improving test outcomes compared to the control group. However, the difference between Treatment 1 and 2 in their response was not that considerable, which implies that the intervention seems to be working.

The coefficients for the school fixed effects, which is where this model diverges from the one estimated previously, show the differences in endline test scores between each of these schools and the reference school, after controlling for other variables. For instance, students in school number 5 scored significantly higher (1.36 points, p=0.022) than the reference school, however this does not seem to be too insightful as quantifying 1.36 points out of 100 is difficult not to attribute to random chance.

The coefficient for the class fixed effect is not statistically significant, suggesting no significant difference in endline test scores between the two classes, after controlling other factors.

OLS Regression Results

Dep. Variable:	test_score	R-squared:	0.228
Model:	OLS	Adj. R-squared:	0.224
Method:	Least Squares	F-statistic:	54.31
Date:	Wed, 19 Mar 2025	Prob (F-statistic):	4.55e-124
Time:	12:11:58	Log-Likelihood:	-9157.6
No. Observations:	2400	AIC:	1.834e+04
Df Residuals:	2386	BIC:	1.842e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	54.6399	0.777	70.278	0.000	53.115	56.164
C(school_id)[T.2.0]	1.5388	0.900	1.710	0.087	-0.226	3.303
C(school_id)[T.3.0]	0.1327	0.900	0.147	0.883	-1.632	1.897
C(school_id)[T.4.0]	2.0359	0.900	2.263	0.024	0.272	3.800
C(school_id)[T.5.0]	2.9684	0.900	3.299	0.001	1.204	4.733
C(school_id)[T.6.0]	-0.4485	0.900	-0.498	0.618	-2.213	1.316
C(school_id)[T.7.0]	0.9343	0.900	1.038	0.299	-0.830	2.699
C(school_id)[T.8.0]	-0.1246	0.900	-0.138	0.890	-1.889	1.640
C(class_id)[T.2.0]	-0.6411	0.450	-1.425	0.154	-1.523	0.241
T1	-0.4778	0.804	-0.594	0.552	-2.054	1.099
T2	-0.3505	0.761	-0.461	0.645	-1.843	1.142
post_binary	1.3013	0.632	2.059	0.040	0.062	2.541
post_binary:T1	12.0382	1.137	10.590	0.000	9.809	14.267
post_binary:T2	13.6128	1.076	12.648	0.000	11.502	15.723

The estimated average test score for the control group in the pre-treatment period is approximately 54.64. The coefficient of 1.3013 (p=0.040) suggests that, on average, the test scores for the control group increased by about 1.30 points from the pre-treatment to the post-treatment period, after controlling for school and class.

The coefficient of 12.0382 (p<0.001) is the DiD estimate for Treatment 1. It indicates that Treatment Group 1 experienced an additional increase of approximately 12.04 points in their test scores from the pre-treatment to the post-treatment period, compared to the control group. The coefficient of 13.6128 (p<0.001) is the DiD estimate for Treatment 2. It indicates that Treatment Group 2 experienced an additional increase of approximately 13.61 points in their test scores from the pre-treatment to the post-treatment period, compared to the control group, this again is different from the first test, where the coefficient was negative for T2. Similar to the previous DiD model, some school fixed effects are statistically significant, indicating differences in average test scores across schools. The class fixed effect is not significant at the 0.05 level.

The estimated treatment effects for both T1 (12.04) and T2 (13.61) from this DiD model are consistent with the coefficients obtained from the second ANCOVA model (T1: 12.04, T2: 13.61), which controlled for school and class fixed effects. This consistency across different methodologies (ANCOVA and DiD) when properly specified strengthens the confidence in these findings.

The results from this second DiD model now show positive and significant effects for both treatments, which contrasts with the negative effect observed for Treatment 2 in the first ANCOVA DiD attempts. This suggests that controlling for school and class differences is crucial for accurately estimating the treatment effects in this context.

Part C

Analysis on how stratification changed the results for the two regressions.

The initial round of analyses, consisting of an ANCOVA model without explicit controls for school and class and the DiD regression, presented a mixed picture of the treatment effects. Both models indicated a positive impact of Treatment 1 on test scores. However, a surprising finding was the statistically significant negative impact of Treatment 2 in both analyses. The first ANCOVA suggested that Treatment 2 led to lower endline scores, while the DiD results suported these findings. These results were counterintuitive and raised questions about the effectiveness of Treatment 2.

The second round of experiments involved analyzing data from a stratified sample, with a focus on controlling for school and class level differences. The second ANCOVA model, which included fixed effects for schools and classes, revealed a significantly different outcome. Both Treatment 1 and Treatment 2 showed substantial positive impacts on endline test scores, with Treatment 2 exhibiting a slightly larger effect than Treatment 1. Similarly, the second DiD regression also demonstrated positive and significant additional increases in test scores for both Treatment 1 and Treatment 2 over time, relative to the control group.

The implementation of stratification clearly played a crucial role in changing the observed results, particularly for Treatment 2. The negative impact seen in the initial analyses was completely reversed to a strong positive impact in the analyses of the stratified data. Furthermore, the magnitude of the positive effect for Treatment 1 also increased. This indicates that the initial findings were likely confounded by factors related to the structure of the data that were not adequately addressed without stratification and the inclusion of school and class controls.

The stratification strategy, which aimed to ensure a more balanced distribution of students with different characteristics across the treatment groups. By subsequently controlling for school and class fixed effects in the ANCOVA and DiD models, the analyses were able to account for pre-existing differences in student performance or other contextual factors at these levels. The reversal of the Treatment 2 effect suggests that, in the non-stratified analyses, Treatment 2 might have been inadvertently associated with lower-performing schools or classes, leading to spurious negative correlation. Stratification and the inclusion of fixed effects helped to isolate the true effect of the treatment by accounting for this underlying structure and potential imbalances.