Assessing the Effect of Class Type on Math Scores of the First Graders

1. Introduction

The Student/Teacher Achievement Ratio (STAR) project was funded by the Tennessee General Assembly and conducted in the late 1980s (Word, E.R., 1990). Based on the literature review, the main characteristics of the project design are stated as below:

- The STAR project had three different level of class type as treatments that including small class, regular class, and regular-with-aide class, which started as the students entering school in kindergarten through the third grade (Achilles et al, 2008).
- There were two separate and independent randomizations happened to arrange experimental units including randomizing teachers to different class types and randomizing students to different classes/teachers (Achilles et al, 2008).
- The schools enrolled in the project had at least one class of each type for proper randomization with a sufficient number of students (Achilles et al, 2008).

Under the randomization, the experiment design is unbalanced randomized complete block design. The primary question of interest is to investigate the effects of class type on teacher-level math scaled scores for the first graders.

2. Exploratory Data Analysis

The original data of 1st grade math scores were explored after removing all the missing data by list-wise deletion method, resulting in 6598 observations with full information of class type, teacher ID and school ID. No extreme values and outliers were removed based on summary statistics. In order to make causal inference on the effect of class type on math score, Stable Unit Treatment Value Assumption (SUTVA) has to hold. Therefore, teacher (or class) was used as the unit of analysis instead of individual student, to avoid the interaction among students within one class. In each school, the median and mean scores for the same class were very close. Since we are more interested in average behavior of math learning skill among students, which is a key quality that shows teacher's performance, mean math scores of students taught by each teacher were calculated for further data analysis.

As a result, 339 mean math scores of each teacher were obtained, with 3 class types in 76 different schools. The distribution of mean math score is nearly symmetric. The data obtained was unbalanced as each school had different number of each class type. To be more specific, some school have only 1 small class, 1 regular class and 1 regular class with aide. This situation might cause unequal variances in data analysis. This issue will be discussed in the following.

3. Two-way ANOVA Model

3.1 Model Building

The design of experiment is a randomized complete block design, and the randomization is to assign teachers to different types of class. In order to alleviate the nuisance effect of different schools on math scores of first graders, we consider different schools (School ID) as a block factor. In this case, we mainly focus on the primary effect of class type on the teacher-level math score. We do not need to analyze the effect on one specific school and there is not enough statistical power to estimate the interaction effect of school and class type, so two-way ANOVA model without interaction terms is used to analyze. In this two-way ANOVA model, we take 3 class types and 76 schools as two factors. The school is called block factor since each school can be seen as one stratum of our experiment. The unit of analysis is the teacher or class, so we use the average math score of the students taught by each teacher as response variables. We set the model notations as below:

- Response variable Y_{ijk}: the average math scaled scores of students taught teacher k in school j of class type i;
- Class type i has 3 levels: $i = 1$ if small class; $i = 2$ if regular class; $i = 3$ if regular class with aide;
- School indicator j has 76 levels, which represents different schools.

The model equation is:

$$Y_{ijk} = \mu_0 + \mu_i + \gamma_j + \epsilon_{ijk}$$

where μ_0 is the overall average math score, μ_i is the effect due to the class type i and γ_j is the effect due to the school j. In the two-way ANOVA model, we assume that the error terms ϵ_{ijk} $N(0, \sigma^2)$: (1). Normality of error terms; (2). Constant Variance of error terms; (3). Independence of error terms.

3.2 Fitting result

We use R to fit two-way ANOVA model without interaction terms. Taking small class (class type 1) as reference group, the results in Table 1 show that the regular class without aide and regular class with aide both tend to have lower average math score. In addition, the mean math score in regular class with aide seems to be similar or slightly higher than the regular class without aide.

Table 1 Fitting Result of Two-Way ANOVA Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	502.5	9.69	51.87	<0.05
class2	-13.37	2.19	-6.1	<0.05
class3	-11.4	2.28	-5	<0.05

Table 2 ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Class type	2	11617.28	5808.64	20.99	<0.05
School	75	136832.53	1824.43	6.59	<0.05
Residuals	261	72224.52	276.72		

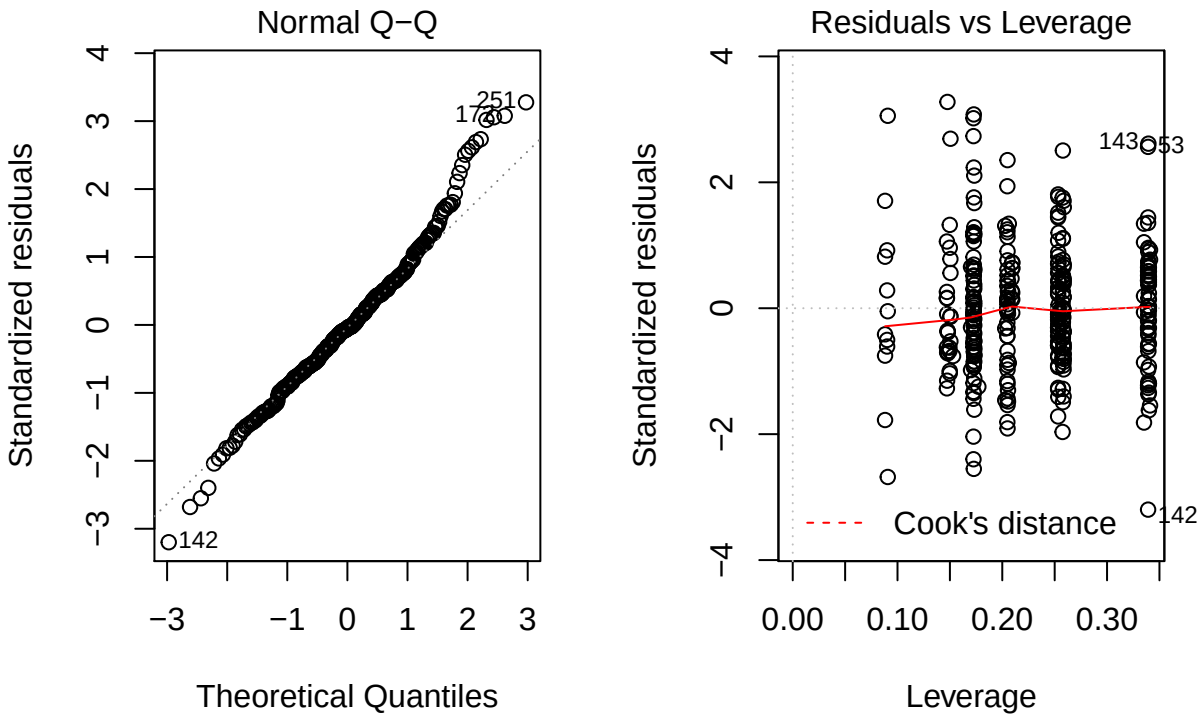
And then we can analyze the results of ANOVA table in Table 2. Given the significant level 0.05, we can conclude that both school and class type are statistically significant. It seems to show that changing the class type or school will impact significantly on average math score. However, the F-test is based on normality assumption of the model so model diagnostics is needed to analyze futher.

3.3 Model Diagnostics

3.3.1 Normality

The assumptions of two-way ANOVA are just the same as One-way ANOVA. Therefore We need to test whether the residuals are normally distributed.

Figure 1 Plot Test for Model Diagnostics



From the Q-Q plot (Figure 1 left) we can see that the most of points are on the diagonal line. But there are still some points that deviate from the straight line in both tails. The distribution of error terms seems to be heavy-tailed, which means it violates the normality assumption.

3.3.2 Homogeneity of Variance

Since the plot test is not obvious to distinguish the pattern, additional statistical test is used to test whether the variance of error terms is constant. Due to the non-normality of error terms, Levene's test is utilized to analyze. From the result of Levene's test (Table 3), given the significant level 0.05, we can see the P-value is less than 0.05. Therefore, the variances of error terms are not equal.

Table 3 Levene's Test

	Df	F value	Pr(>F)
Group	223	2.1	<0.05
Residuals	115		

3.3.3 Independence and Outliers

The independence of the error terms is due to double randomization processes. Within each school block, both teachers and students were randomly assigned to a different type of class, which eliminate the correlations caused by the adjacent experimental units in time and space.

From the Residuals vs Leverage plot (Figure 1 right), we can see there is no obvious extreme value or outliers. Therefore, we believe the model we build is not affected by outliers or high-influential points.

4. Statistical Inference

4.1 Scheirer–Ray–Hare Test

According to the diagnostics of two-way ANOVA model, the distribution of residuals are heavy-tailed with non-constant variance across group. In this case, we cannot use F-test or other relative normality-based statistical tests to analyze the effect of class type on math scaled scores across teachers. However, the Scheirer–Ray–Hare(SRH) test, which is an extension of the Kruskal-Wallis test, can be used to test whether the math scaled scores is affected by class types and different schools.

The SRH test is a nonparametric test which is suitable to randomized complete block design. In this case, we are not interested in the effect of class type on math scores in one specific school. Even though the conclusion and interpretation of interaction terms in SRH test is conservative, we can still use this test to examine the main effect of class type. The null hypothesis and alternative hypothesis are listed below:

H_0: All types of class have the same average math scaled scores ($\mu_1 = \mu_2 = \mu_3$).

H_a: Not all types of class have the same average math scaled scores (at least one $\mu_i \neq \mu_j$).

Table 4 SRH Test

	Degree of Freedom	Sum of Square	H-Statistic	P-value
class	2	164742.2	17.15	0
sid	75	2042131.6	212.61	0
class:sid	146	598192.3	62.28	1
Residuals	115	441421.8		

From the result of SRH test (Table 4), given the significant level 0.05, the null hypothesis of equal sample mean is rejected significantly. It means that the type of class can affect the math scaled scores in the first grade, the effect of class type on scores is significantly different across teachers.

4.2 Multiple Comparison

Since the family-wise difference of average math scores shows significance, Dunn's Test can be used to pinpoint which specific class type tends to have higher average math score. In this case, Dunn's multiple comparison test, which is also a nonparametric test, can analyze the pair-wise difference among different types of class. In order to maintain Type I error (falsely reject null hypothesis), Benjamini-Hochberg Procedure is applied to adjust the p-value. For each comparison, the null hypothesis and alternative hypothesis is that:

H_0: The average math scaled scores is the same between two types of class ($\mu_i = \mu_j$).

H_a: The average math scaled scores is not the same between two types of class ($\mu_i \neq \mu_j$).

Table 5 Dunn's Multiple Comparison Test

Comparison	Z-Statistic	P-value	Adjusted p-value
Small Class - Regular Class	4.10	0.00	0.00
Small Class - Regular Class with aide	2.41	0.02	0.02
Regular Class - Regular Class with aide	-1.51	0.13	0.13

According to the result of Dunn's multiple comparison test (Table 5), given the significant level 0.05, the smaller class tends to have higher average math scores in the first grade compared with regular class with or without aide. In addition, since the adjusted p-value for comparison of regular class with aide and regular class without aide is 0.13 (>0.05), these two types of class tends to have similar average math scores across teachers.

5. Conclusion and Disccusion

5.1 Causal Inference

As we analyze above, the smaller class tends to have higher math scores. In order to analyze the causal effect of class type on average math score, we need to check the assumptions of causal inference: (1). Causal ordering: in this project, the math score does not influence the randomization of class type; (2). No Spillover effect: In the randomized block design, one teacher only teaches one specific class and one is not influenced by other teachers; (3) Same version of treatment: all teachers are teaching one type of class and there is no teacher teaching other class type than regular or small ones; (4) The stable unit treatment value assumption: Through the randomized complete block design, we can control the variability caused by the school type and reduce the error within each school (as a blocking factor), which helps to satisfy SUTVA; (5) Positivity assumption: Randomization eliminates the conditioning on other variables and ensures the probability of teacher-level unit receiving treatment is positive since there is no teacher who is not teaching in a class.

In conclusion, when we set teacher as a unit, we can make a causal statement more validly comparing with project 1. The class type shows causal effect on math score across teachers. However, there are many other possible factors which could affect the math score such as the social-economic status or free-lunch status. In the further investigation, more advanced and complicated methods may be adopted to analyze our data.

5.2 Discussion

In project 1, we find different class types have different teaching qualities. The smaller class tends to have the highest average math score and the regular class without aide tends to have the lowest math score compared with others. However, the result of project 2 shows that the difference of scores between a regular class and regular class with an aide is not significant, which means that after considering school ID as blocking factor, the accuracy of the model has been improved. It helps reduce the error and enhance our statistical power since the block factor helps to explain a part of variance of response variables. In addition, Both projects found a significant effect of class size on 1st grade math score scores no matter it's on student level or teacher level.

The biggest difference between the two projects is different randomization approaches. Project 1 is a completely randomized experiment while project 2 is a stratified randomized experiment. Stratification rules out the nuisance factors that affects the outcomes of the response variable (Imbens & Rubin, 2015). Therefore, project 2 can estimate the effects of class types without the influence of the different types of schools that may cause different results between the two projects.

Reference

Achilles, C. M., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., ... & Word, E. (2008). Tennessee's Student Teacher Achievement Ratio (STAR) roject'. URL: http://hdl. handle. net/1902.1/10766.

Imbens, G. W., & Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

Word, E. R. (1990). The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report (1985-1990)