The effects of class types on the 1st grade math score of teacher

1/30/2020

1 Introduction

Large literature on the effect of school resources on student achievement are generally ambiguous and conflicting; even quantitative summaries cannot reach consistent conclusion (Krueger 1999). In this project, we focus on evaluating the effects of class type and school on test scores based on the Tennessee Student/Teacher Achievement Ratio (STAR) project, which is a four-year longitudinal class-size study including more than 7,000 students in 79 schools. The STAR project data is available from Harvard Dataverse online library. (Achilles et al. 2008). In the experiment, within each school, a completely randomized experiment is conducted: students and teachers were randomly assigned into one of three class types: small class (13-17 students per teacher), regular class (22-25 students per teacher), and regular classes with a teacher's aide. To be eligible for this experiment, a school had to have sufficient number of students to form at least one class of each of the three types (Imbens and Rubin 2015).

The purpose of this analysis is to explore simultaneously the effects of class type and school based on the first-grade math scores from STAR data with teacher as the unit through a two-way analysis of variance (ANOVA) model. Among all variables 379 variables from the full `STAR` dataset, we only explore variables that may have influence on 1st grade math score of teachers, which are teachers' gender, ethnicity, class type in 1st grade (g1classtype), highest degree of 1st grade teacher (g1thighdegree), teacher's career ladder level in 1st grade (g1tcareer), and years of teacher's total teaching experience in 1st grade (g1tyears). As shown in Table 4 in the appendix, the composition of teachers' gender and ethnicity are relative stable across all three class types. Nearly all teachers are female with only 1.5% male in the small class, and White dominates, followed by Black, together accounting for all teachers in each class type. The proportion of missing values of 1st grade math score in each class type is around 3%-4%. 1st grade teachers' career ladder and years of teaching experience are slightly different across class types. For examples, the ratios of probation teachers are around 6.6% for regular + aide class and small class types, however, 14.8% for regular class type. The average teaching experience of teachers are around 12 years for small and regular + aide class types, but 10 years in small class type. Generally, we could see that the random assignment of teachers to classes of different types is valid, and the resulting inferences are valid for the effect on the teachers of being assigned to a particular type of class.

To better indicate the general performance of each teacher, we focus on the median scores on a mathematics test over all students for their teacher. We choose median instead of mean value to avoid the effects from outliers. In total, there are 76 schools and 339 teachers. All schools have small and regular class types, but 4 schools do not have regular-with-aide class type. We discard schools that do not have at least one class of each of the three types, leaving us with 72 schools, which creates 72 strata. The total number of teachers in this reduced data set is N = 325. Out of these 325 teachers, N1 = 118 are assigned to small classes, N2 = 107 and N3 = 100 are assigned to regular classes and regular-with-aide classes respectively. Figure 3 in the appendix shows the distribution of teachers' performance, which is the median math scores of their 1st grade students, in different class types.

The results indicate that both class size and school does have a statistically significant effect on teachers' performance in terms of median of test scores over all his/her students. More precisely, teachers who were assigned to small classes have statistically significant higher score than those who were assigned to regular or regular-with-aide classes, while we do not find significant difference between the math scores of teachers from regular and regular-with-aide classes.

2. Analysis

2.1 Two-way ANOVA Model

To study the effects of class types on math scaled scores in 1st grade with the school indicator as the other factor, we will use a two-way ANOVA model. Our general model equation is shown as following:

$Y_{i,j,k} = \mu_{i,j} + \alpha_i + \beta_j + \epsilon_{i,j,k}$ where $i = 1, ..., l$; $j = 1, ..., m$; $k = 1, ..., n_i$.

The assumptions of our model:

- All the subjects ($Y_{i,j,k}$) are randomly sampled (independent).
- All levels of factor i are independent
- All levels of factor j are independent
- The errors $\epsilon_{i,j,k}$s are normally and independently distributed with mean 0 and variance σ^2.

Explanation of the natations:

- The dependent variable $Y_{i,j,k}$ represents the k^{th} observation with treatment i and j. In terms of our problem, factor i represents different class types, and factor j represents different schools. As an illustration, $Y_{1,2,3}$ would represent the 3rd observation in class type 1 of the 2nd school.
- $\mu_{i,j}$ is the common effect (overall mean) for each class type of each school (as $i = 1, 2, 3$ represent small, regular and regular-with-aide class type respectively, and $j = 1, 2, ..., 64$ represents 64 different schools in the STAR dataset).
- $\epsilon_{i,j,k}$ represents the random indiviual error in the k^{th} observation on the treatment i and j., which are the unobserved random variables in this experiment.

We do not include the interaction term in our two-way ANOVA model since the interaction term reflects the class type influences within a single school, which is not the interest of this analysis. But rather, we care about hypotheses and treatment effects across all schools. Moreover, the sample sizes are not big enough so that we cannot obtain precise estimates of the class type effects within any one school.

2.2 Model Diagnostic

After fitting our model, it is necessary to use some diagnostic plots and tests to check whether data has violated the model assumptions of ANOVA mentioned above. We start with looking for outliers in our raw dataset based on the "Semi-Studentized Residuals Method" and "Studentized Residuals Method" since outliers, We test the independence of subjects and factors. Then, we use the "Residuals vs Fitted plot" with the "Levene Test" to check the homogeneity of variance. Lastly, the histgram of residuals and QQ-plot are used to test the normality of residuals.

2.3 Pairwise Comparison

In two-way ANOVA test, a significant p-value indicates that some of the group means are different, but we don't know which pairs of groups are different. Therefore, we could conduct pairwise-comparison to find out which group means are significantly different. In this section, we consider Tukey's procedure, Bonferroni's procedure, and Scheffe's procedure. Since we care about all pair-wise comparisons, Tukey's procedure is the best among all three procedures. And here we show the Tukey's procedure pair-wise comparison results.

Since we define teachers' performance as the median math score of the students, for all three pair-wise comparisons, we have three sets of hypothesis and tests:

small-regular:

H0: the teachers' performance mean between small and regular classes are equal;

H1: the teachers' performance mean of small and regular classes are not equal.

regular+aide-regular:

H0: the teachers' performance mean between regular with aide and regular classes are equal;

H1: the teachers' performance mean of regular with aide and regular classes are not equal.

regular+aide-small:

H0: the teachers' performance mean between regular with aide and small classes are equal;

H1: the teachers' performance mean of regular with aide and small classes are not equal.

Decision rule: reject H0 at significant level α if the p-value of the test $< \alpha$.

3. Results

3.1 Fit two-way ANOVA model

The result of the Two-Way ANOVA is illustrated following:

Model: 1st grade math score ~ class type + school

Table 1: Summary of the Two-Way ANOVA Model.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
class	2	11974	5987	19.68	0
school	71	137232	1933	6.36	0
Residuals	251	76345	304	NA	NA

From Table 1, we can observe that both p-value of class type and school are lower than 0.05, so we can conclude that both class type and school are statistically significant. School is the most significant factor variable. These results would lead is to believe that changing class type or school, will impact significantly the mean math score of 1st grade.

Also we can get sum of squared values together with degrees of freedom:

$SSE = 7.634 \times 10^4$, $df = 251$, $MSE = \frac{SSE}{df(SSE)} = 304.163$

$SSTO = 2.256 \times 10^5$, $df = 324$, $MSTO = \frac{SSTO}{df(SSTO)} = 693.004$

$SSTR = 1.492 \times 10^5$, $df = 73$, $MSTR = \frac{SSTR}{df(SSTR)} = 2043.924$

3.2 Model diagnostics and Sensitivity analysis

3.2.1 Checking outliers

After calculating in R using both "Semi-Studentized Residuals" and "Studentized Residuals" approaches, we find that there are no outliers in the raw data. So, we may not remove any sample from our raw dataset.

3.2.2 Checking independence

In this experiment, the study randomly assigned student to different class type and school, also teachers are also randomly assigned to the class. There is also no evidence indicating the association of math score between different students. Thus, we can conclude that all factors and subjects are independent in the dataset.

3.2.3 Checkinig homogenerity of variance

According to the Residuals vs Fitted values plot in Figure 1 in the appendix, there is no evidence showing that the variance across groups are statistically significantly different.

We try to confirm this finding by using a Levene Test. And the results are shown in Table 2.

Table 2 Summary of the Levene's Test for the Two-Way ANOVA Model.

Source of variation	Drgrees of freedom	F value	Pr(>F)
group	215	1.88	0.00015
	109		

From the result of Levene's test above, it gives a P-value of 7.1e-06 which is much lower than the siginificant level of 0.05.This indicates an issue of unequal variance, which conflicts to our findings in the Residuals and fitted plot.We believe the major reason why this test has not passed is the small sample size in our dataset. Since we only have 325 observations but 216 cells, the number of observations in each cell is very limited, which makes the test hard to pass.

3.2.4 Checking Normality

According to the histogram of math score in Figure 1 in the appendix, the bell-shaped distribution of math score can be observed. From the QQ-plot in Figure 1 in the appendix, we can assume normality of the data as most of the points falls approximately along the reference line.

3.3 Pair-wise comparison

The Tukey's procedure calculated class type pair-wise comparison results are shown in the Table 3 below. The confidence intervals for class type pair-wise comparison are shown in Figure 2 in the appendix.

Table 3: Class type pair-wise comparison confidence interval

	diff	lwr	upr	p adj
REGULAR CLASS-SMALL CLASS	-13.014	-18.50	-7.53	0.000
REGULAR + AIDE CLASS-SMALL CLASS	-12.162	-17.75	-6.57	0.000
REGULAR + AIDE CLASS-REGULAR CLASS	0.852	-4.87	6.57	0.934

Since p-value of the pair-wise comparison between regular-small and regular+aide-small are much smaller than 0.005, we reject the null hypotheses of regular-small and regular+aide-small and conclude that the mean teachers' performance for regular-small and regular+aide-small are different from the others with at least significant level 99.5%. For regular+aide-regular comparison, p-value is greater than 0.5, therefore, we fail to reject the null hypothesis and conclude that we are not able to find the teachers' performance difference between regular+aide class and regular class.

4.0 Discussion

Using a two-way ANOVA model, we find that both class size and school have effect on teachers' performance. Among all 1st grade teachers, those randomly assigned to the small class size outperform those teachers in both the regular and regular + aide groups in terms of the median math scores of the students. Teachers in small classes show 13 points increase on average in their median math scaled scores compared with teachers in regular classes, and an average 12 points higher than those in regular-with-aide classes. We find no evidence that aide could help improve teachers' outcomes in regular classes (no significant difference between regular and regular-with-aide classes).

We can safely make a causal statement between class types and teacher's performance (median math score of each teacher) since all the following assumptions are valid:

- **Assumption 1: Stable Unit Treatment Value Assumption (SUTVA)**, which means 1) the assignment of treatment to one person does not affect the potential outcomes of others, and 2) treatments are stable. In our case, one teacher's potential performance does not depend on the class type assigned to other teachers, and teachers' potential performance will not be affected by whether teachers from different class types share or discuss teaching materials (no interference). Moreover, given the structure of the experiment, we can make sure that class types are defined under the same criteria across all schools (treatment are stable). Therefore, the SUTVA assumption holds in this case.
- **Assumption 2: Ignorability**, which means the assignment of treatment is independent of the potential outcome. STAR experiment is a stratified randomized experiment, where teachers and students are grouped together into schools in advance naturally. Within each school, a completely randomized experiment is conducted, making sure that other variables, such as teachers' experiment or education level, are relatively the same across classes of different types (as shown in Table 4), and only source of difference in teachers' performance is from the assignment of class type itself.

Recall the results of project 1, we cannot conclude that there are causal relations between class types and mean math scores of students since we do not know about the interference among students in different classes. For example, students from different class types may review class materials together after before exam, and the performance of their math tests will be influenced by these study groups. However, in project 2, we can make the causal statement since we change the unit of analysis to the teacher from the individual student, which help justify the no-interference part of SUTVA. Irrelevant of the interference among students, the resulting inferences are valid for the effect on the teachers of being assigned to a particular type of class.

Reference

Achilles, C.M., Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. 2008. "Tennessee's Student Teacher Achievement Ratio (STAR) project." Harvard Dataverse. doi:10.7910/DVN/SIWH9F.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics* 114 (2). MIT Press: 497–532.

Appendix

Table 4 Summary of the STAR data.

	SMALL CLASS (n=1925)	REGULAR CLASS (n=2584)	REGULAR + AIDE CLASS (n=2320)	Overall (n=11601)
teacher gender				
MALE	29 (1.5%)	0 (0%)	0 (0%)	29 (0.2%)
FEMALE	1877 (97.5%)	2584 (100%)	2320 (100%)	6781 (58.5%)
Missing	19 (1.0%)	0 (0%)	0 (0%)	4791 (41.3%)
teacher ethnicity				
WHITE	1564 (81.2%)	2166 (83.8%)	1893 (81.6%)	5623 (48.5%)
BLACK	342 (17.8%)	418 (16.2%)	427 (18.4%)	1187 (10.2%)
ASIAN	0 (0%)	0 (0%)	0 (0%)	0 (0%)
HISPANIC	0 (0%)	0 (0%)	0 (0%)	0 (0%)
NATIVE AMERICAN	0 (0%)	0 (0%)	0 (0%)	0 (0%)
OTHER	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Missing	19 (1.0%)	0 (0%)	0 (0%)	4791 (41.3%)
scaled math score				
Mean (SD)	539 (44.1)	525 (41.7)	530 (42.9)	531 (43.1)
Median [Min, Max]	535 [425, 676]	523 [408, 676]	529 [404, 676]	529 [404, 676]
Missing	58 (3.0%)	77 (3.0%)	96 (4.1%)	5003 (43.1%)
teacher degree				
ASSOCIATES	0 (0%)	0 (0%)	0 (0%)	0 (0%)
BACHELORS	1224 (63.6%)	1793 (69.4%)	1439 (62.0%)	4456 (38.4%)
MASTERS	665 (34.5%)	791 (30.6%)	838 (36.1%)	2294 (19.8%)
MASTER +	0 (0%)	0 (0%)	0 (0%)	0 (0%)
SPECIALIST	17 (0.9%)	0 (0%)	21 (0.9%)	38 (0.3%)
DOCTORAL	0 (0%)	0 (0%)	22 (0.9%)	22 (0.2%)
Missing	19 (1.0%)	0 (0%)	0 (0%)	4791 (41.3%)
g1tcareer				
Not on Ladder	125 (6.5%)	201 (7.8%)	180 (7.8%)	506 (4.4%)
APPRENTICE	203 (10.5%)	254 (9.8%)	261 (11.2%)	718 (6.2%)
PROBATION	128 (6.6%)	382 (14.8%)	156 (6.7%)	666 (5.7%)
LADDER LEVEL 1	1309 (68.0%)	1635 (63.3%)	1548 (66.7%)	4492 (38.7%)
LADDER LEVEL 2	15 (0.8%)	0 (0%)	99 (4.3%)	114 (1.0%)
LADDER LEVEL 3	126 (6.5%)	89 (3.4%)	76 (3.3%)	291 (2.5%)
PENDING	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Missing	19 (1.0%)	23 (0.9%)	0 (0%)	4814 (41.5%)
teacher experience (years)				
Mean (SD)	12.2 (8.65)	10.3 (8.70)	12.7 (9.24)	11.6 (8.94)
Median [Min, Max]	12.0 [0.00, 39.0]	8.00 [0.00, 42.0]	11.0 [0.00, 39.0]	10.0 [0.00, 42.0]
Missing	19 (1.0%)	0 (0%)	0 (0%)	4791 (41.3%)

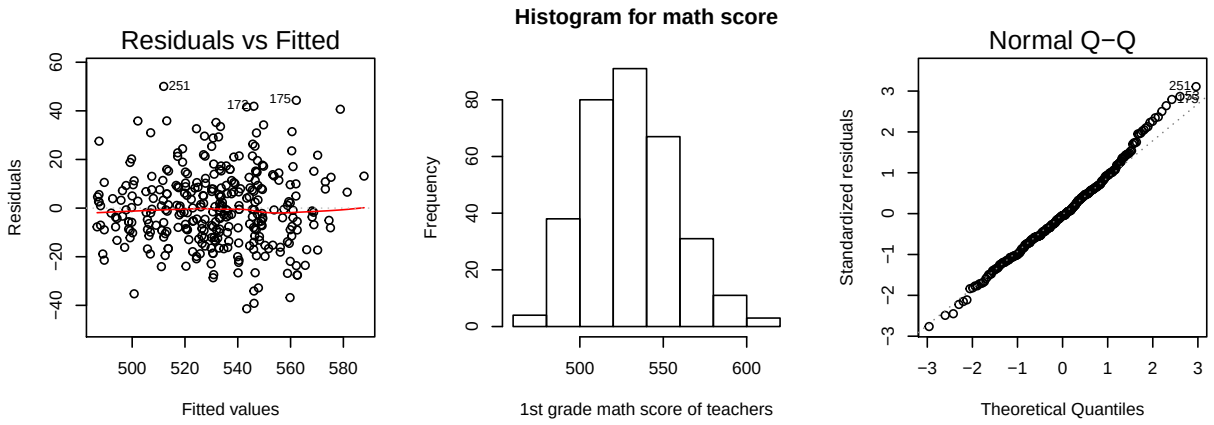

Figure 1: Model Diagnostic Figures

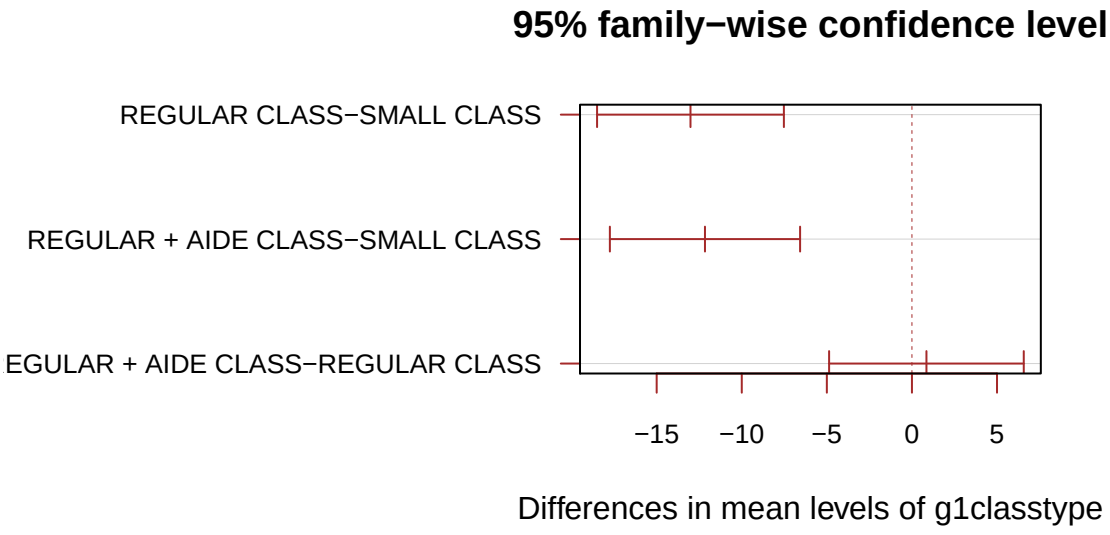

Figure 2: Class type pair-wise confidence interval

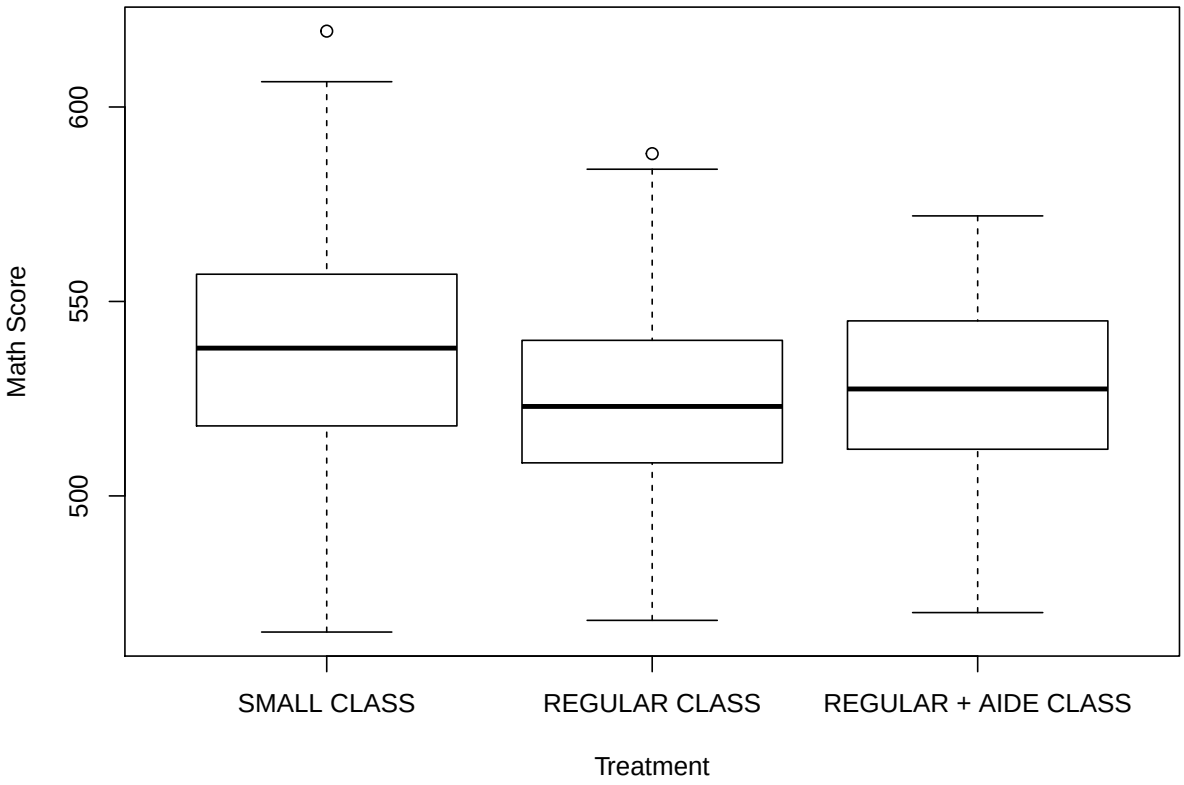

Figure 3: Math score distributions for different class types