STA 207

Effects of Class Size on Math Teacher Performance in Tennessee

1-31-2020

1.0 Introduction

1.1 Background

The Student/Teacher Achievement Ratio (STAR) was a four-year longitudinal class-size study funded by the Tennessee General Assembly. Over 7,000 students in 76 schools were randomly assigned into one of three interventions: small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students with a full-time teacher's aide). The interventions were consistent between grades kindergarten through third.

Besides following standard procedures that ensured confidentiality and ethics in human subjects' research, Project STAR also highlights the important features of:

1. *Each school included in the study had to have a large enough student body to form as least one of each of the three class types.*
2. *Students and teachers were randomly assigned to their class type.*

Project STAR is an example of **stratified randomized design**, where experimental units are grouped together according to certain pre-treatment characteristics into strata. Within each stratum, a completely randomized experiment is conducted. In the case that there exist population structures that associate or covary with the experimental outcome, stratified randomized experiments generally are more informative than completely randomized experiments (Suresh K., 2011). The goal of a stratified study is to identify treatment effects across all strata.

Because it is reasonable to expect systemic differences in educational outcomes across schools, due to reasons such as demographics, a specific school could become a confounding variable and influence the outcome of the Project STAR research. Therefore, each school should be viewed as a stratum in analyzing Project STAR data.

To expand on previous findings, teachers, rather than individual students, will be our experimental units. The adjustment of unit will facilitate the making of a causal statement as to the effect of class size on educational outcome, because teacher experimental units more convincingly satisfy the SUTVA and independence assumptions for causal inferences.

1.2 Questions of Interest

1. Is there a significant difference in a first-grade teacher's teaching performance in math across the class sizes?
2. Are teachers' performances relatively stable between different schools?
3. Does our ANOVA model fit well with the data? In other words, are the analysis of variance assumptions satisfied?
4. Can we draw a causal conclusion that class sizes affect the class average math scores of first-grade teachers?

2.0 Analysis Plan

2.1 Population and study design

A two-way ANOVA test is fitting for answering our questions of interest under the stratified randomized design. One factor in the ANOVA model will be class size, whose main effect is of primary interest in this study. The other factor will be school ID, in order to control for and observe the stratum effect. Our model is appropriate to answer the questions of interest because it captures the effect of each treatment on the class' median math performance while controlling for other external factors by blocking by school ID.

In order to treat teachers as the experimental units, we will use the *median* scaled 1st grade math score of all students under each teacher for our analysis. The median score of a class better reflects the class' performance. In addition, the median is usually a more robust summary statistic than the mean, because it is less affected by outliers.

2.2 Statistical Analysis

2.2.1 Descriptive Analysis

We will examine the following aspects to characterize the independent and response variables:

- Completeness and the level of balance across the strata;
- Distribution of response variable, median scaled 1st grade math score, across schools (strata);
- Distribution of response variable across class sizes;

2.2.2 Main Analysis

For our main analysis, we will construct the following factor effects ANOVA model for the classroom median math score:

$$Y_{ijk} = \mu_{..} + \tau_i + \beta_j + \varepsilon_{ijk}$$

for $i \in [1,2,3,4]$, $j \in [1,2,\dots,76]$, and $k \in [1,2,\dots,338]$

Our model is constrained such that $\sum_{i=1}^{4} \tau_i = 0$ and $\sum_{i=1}^{76} \beta_j = 0$, which yields the following interpretations.

- $\mu_{..}$ represents the overall classroom median score across all treatment levels.
- τ_i represents the effect of each class size on the overall median math score.

- β_j represents the effect of each school on the overall median math score.

The proposed model does not consider the interaction effect between class type and school ID. We made this decision for two reasons. First, the inclusion of an interaction term would increase the number of parameters needing estimation to a number that exceeds the sample size. This would impede us from fitting an accurate model. The second reason is that of relevance. The purpose of the analysis is to examine the main effects of class size on a class' median scaled math score. If we include an interaction term, our model now describes the effect of class sizes within individual schools. Due to both of these two arguments, we will omit the interaction parameter between class size and school ID from our model.

2.2.3 Hypothesis Testing

Due to the unbalanced design of the experiment, the factor effect component of the sum of squares are no longer orthogonal. Therefore, we will use the general linear F-test as the mechanism to test if each specific component can be dropped from the model. We will compare the SSE under the full model with the SSE under the reduced model using the F statistic:

(1) $F^* = \frac{\frac{SSE(R)-SSE(F)}{df_R - df_F}}{MSE(F)}$, where SSE(R) is SSE under the reduced model, df_R is the degree of freedom for the reduced model, SSE(F) is SSE under the full model, and df_F is the degree of freedom for the full model. MSE(F) is the MSE of the full model

(2) F^* follows the F distribution, $F_{(df_R - df_F), df_F}$, under the null hypothesis (H_0).

(3) We would reject H_0 at level α if $F^* > F(1-\alpha; (df_R - df_F), df_F)$, or if the p-value $< \alpha = .05$

2.2.3.1 Class Size Effects

Then, we want to test whether class size effects are present:

$$H_0: \tau_1 = \tau_2 = \tau_3 = 0 \quad | \quad H_a\text{: not all } \tau_i{}'s \text{ equal zero.}$$

- Full model: $Y_{ijk} = \mu_{..} + \tau_i + \beta_j + \epsilon_{ijk}$.
- Reduced model is: $Y_{ijk} = \mu_{..} + \beta_j + \epsilon_{ijk}$.

If we reject H_0 at level $\alpha = .05$, we can conclude that the effects of class size are present.

2.2.3.2 Pairwise Class Size Effects

We will also do pairwise comparisons among the three class sizes. The Tukey's procedure will be used as it yields a conservative result when sample sizes are unequal.

First, we define the difference between two factor level means $D_{ii'} = \mu_{i.} - \mu_{i',}$, where $\mu_{i.} = \mu_{..} + \tau_i$. The point estimate for $D_{ii'}$ is $\widehat{D}_{ii'} = \overline{Y_{i..}} - \overline{Y_{i'..}}$. Since $\overline{Y_{i..}}$ and $\overline{Y_{i'..}}$ are independent, the variance of $\widehat{D}_{ii'}$ is $\sigma^2\{\widehat{D}_{ii'}\} = \frac{\sigma^2}{76^2}\sum_j (\frac{1}{n_{ij}} + \frac{1}{n_{i'j}})$, and the estimated variance of $\widehat{D}_{ii'}$ is $s^2\{\widehat{D}_{ii'}\} = \frac{MSE}{76^2}\sum_j (\frac{1}{n_{ij}} + \frac{1}{n_{i'j}})$.

Then, we complete the simultaneous testing using the following null and alternative hypotheses:

$$H_0: D_{ii'} = 0 \ | \ H_a: D_{ii'} \neq 0$$

If we control the family-wise confidence coefficient at level 1-α, the confidence interval for $D_{ii'}$ is of the form:

$$\widehat{D}_{ii'} \pm T \times s(\widehat{D}_{ii'}), \text{ where } T = \frac{1}{\sqrt{2}} q(1-\alpha; 3, n_T - 3 * 76)$$

We can check whether zero is in each interval. If zero is contained, we will not be able to reject the null hypothesis.

2.2.3.3 School Effects

Although the class size effects are of our primary interests, we also want to test whether school effects are present:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_{76} = 0 \quad | \quad H_a\text{: not all } \beta_i{}'s \text{ equal zero.}$$

- Full model: $Y_{ijk} = \mu_{..} + \tau_i + \beta_j + \epsilon_{ijk}$.
- Reduced model is: $Y_{ijk} = \mu_{..} + \tau_i + \epsilon_{ijk}$.

If we reject H_0 at level α, we conclude the effects of a specific school are present.

2.3 Model Diagnostics

We will use a Q-Q plot, a histogram and the Shapiro-Wilk test to inspect the normality of the distribution of the residuals. A residuals-versus-fitted value scatter plot and the Levene test will be used to examine equality of residual variance. Independence of residuals and outlying data points will also be discussed.

3.0 Results

3.1 Descriptive Analysis

After filtering out entries with missing datapoints, our dataset included the median scaled 1st grade math scores from 338 teachers from 76 schools. 72 out of 76 schools had at least one complete set of the class size treatments (Figure 1a). The four schools with incomplete treatment sets all had both regular classes and small classes. Due to the small number of incomplete strata and the presence of more than one treatment within them, we retained these schools in the analysis (Kutner et al., 2005, p.966). The distribution of classroom median scaled math scores differs across class sizes (Figure 1b), with large within-group variance that causes the distributions to overlap with one another. Figure 1c highlights the high variability in teacher performance across distinct schools.

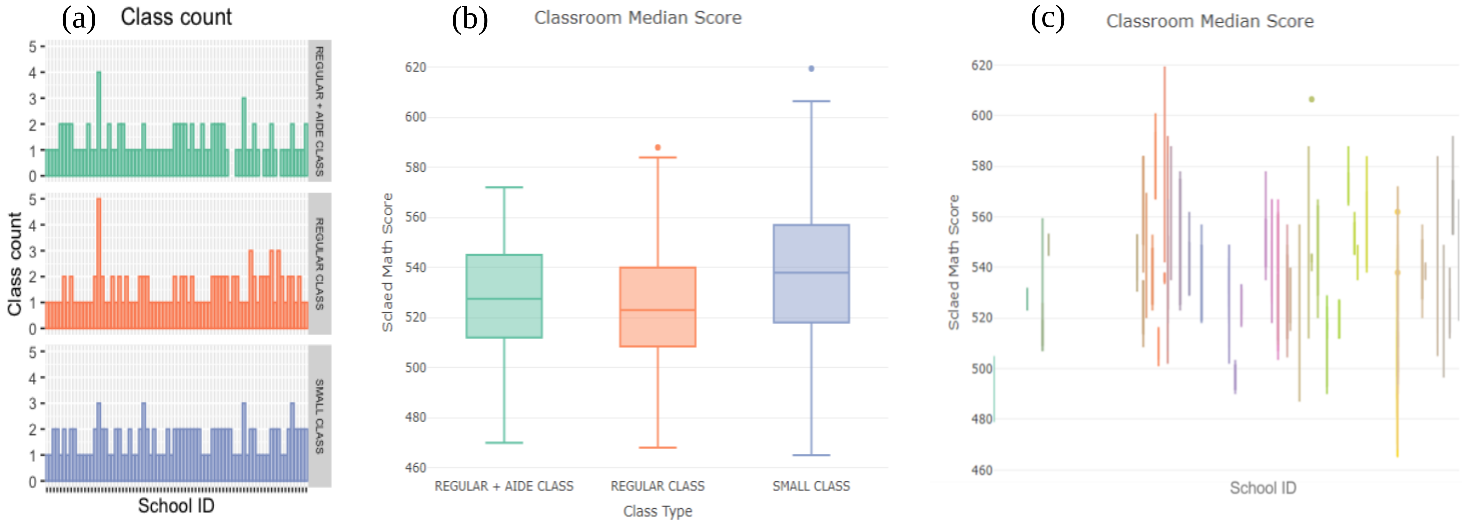

Figure 1. (a). Results of proposed descriptive analysis show completeness of strata. (b, c) Differences in median scaled 1st grade math score across class types (b) and schools (c).

3.2 Main Analysis

Table 1 includes the degrees of freedom, sum of squared error, mean squared error, F-test statistics and corresponding p-values for our ANOVA model. Both p-values for the variables class type and school ID fall below our specified $\alpha = 0.05$ and provide evidence that their inclusion in the model accounts for a significant amount of the variability in the median scaled math scores of teachers. The hypotheses for this test are discussed in section 2.2.3.

Table 1: ANOVA Table for model in 2.2.2

Variance Components	Degrees of Freedom	Sum of Squared	Mean of Squared	F^*	P-value
Class type	2	10796.02	5398.01	17.25	<0.001
School ID	75	149258.51	1990.11	6.36	<0.001
Residuals	260	81345.30	312.87		

3.3 Model Diagnostics

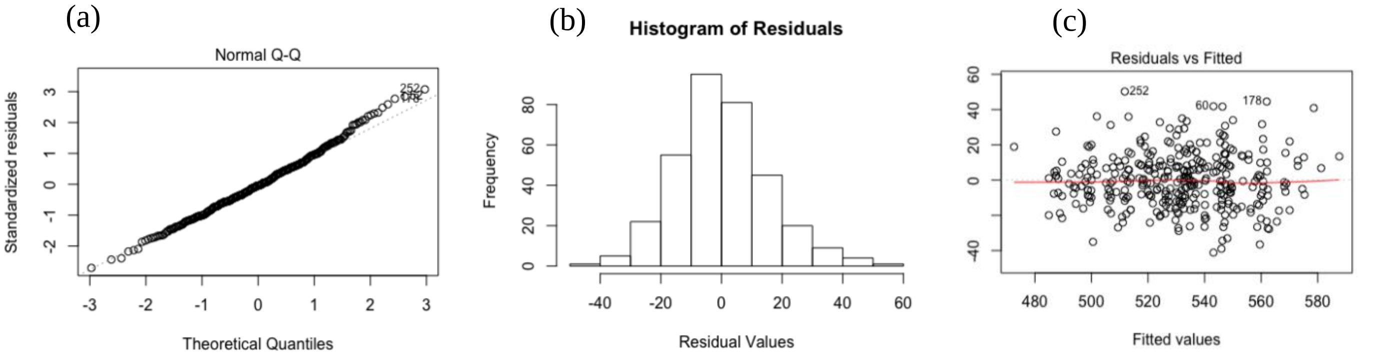

Figure 2: Visual diagnostics of ANOVA model assumptions. (a). Normal Q-Q plot of residuals. (b) Histogram of model residuals. (c) Residual-versus-fitted value scatter plot.

3.3.1 Normality:
From the Q-Q plot of residuals (Figure 2a), we observe that most points lie on the straight line, which is close to what we expect to see from a normal distribution. There are some points at the right tail which have a higher probability mass than expected; however, these deviations do not dominate the plot. Thus, the normality assumption is largely satisfied from the Q-Q plot. From the histogram in Figure 2b, we can observe that the distribution of the residuals looks symmetric about the mean and is bell-shaped, similar to what is seen from a normal distribution. To further test the normality of errors, a Shapiro-Wilk test is used on the distribution of the residuals. This test examines if our data follow a normal distribution.

The null and alternative hypotheses of the Shapiro-Wilk test are:
H_o: The residuals are normally distributed | H_a: The residuals are not normally distributed

The W statistic is 0.99 and the p-value is 0.13, which is greater than the significance level of 0.05. Thus, we fail to reject the null hypothesis, and conclude that there is no evidence that the distribution of the residuals does not follow a normal distribution.

3.3.2 Equal Variances:
While Figure 2c visually satisfies the equal variance assumption, a Levene's test is used to further examine the equal variance assumption for both independent variables.

The null and alternative hypotheses of the Levene's test are:
H_0: The residual variances are equal across groups. | H_a: Not all residual variances are equal across groups.

In this test, both p-values (0.71 for class type and 0.92 for school ID) are greater than the significance level of 0.05. Thus, there is no evidence that residual variances are different across treatment groups, satisfying the equal variance assumption.

3.3.3 Independence:
This experiment is randomized in two ways and represents the best controlled environment to achieve independence. First, teachers are randomly assigned to each class type. Second, every student is randomly assigned to each teacher. While this independence may not hold completely from the time of randomization to the time that the test scores were recorded, the randomization of the experiment and relatively large sample size allow us to assume independence.

3.3.4 Outliers:
Outliers are defined as having a studentized residual value greater than 3 in absolute value. The entry with teacher ID 24475510 was determined an outlier, (z = 3.13). Upon closer inspection, we did not find sufficient evidence to exclude it.

3.4 Hypothesis Testing
We use significance level 0.05 for all the following tests.

3.4.1 Class Size Effects
The results of the F-test for class type main effects are shown in Table 2.

Table 2: Test for Factor Main Effects

Model	Degree of Freedom	SSE	F^*	P-value
Full	334	221371		
Reduced	336	232391	8.3137	0.0002995

Since the p-value = 0.0002995, we reject $H_0: \tau_1 = \tau_2 = \tau_3 = 0$ and conclude that there are class type main effects.

3.4.2 Pairwise Class Size Effects

Pairwise comparisons of factor level means
From Figure 3, we could see that zero is only contained within the first confidence interval ("Regular – Regular with Aide"), not within the second or third confidence intervals. We therefore conclude that, at a family-wise level $\alpha = 0.05$, the means of small classes and regular classes with aides, and the means of small classes and regular classes are different, while there is no significant difference between regular classes and regular classes with aides. Small classes outperformed both regular classes and regular classes with aides.

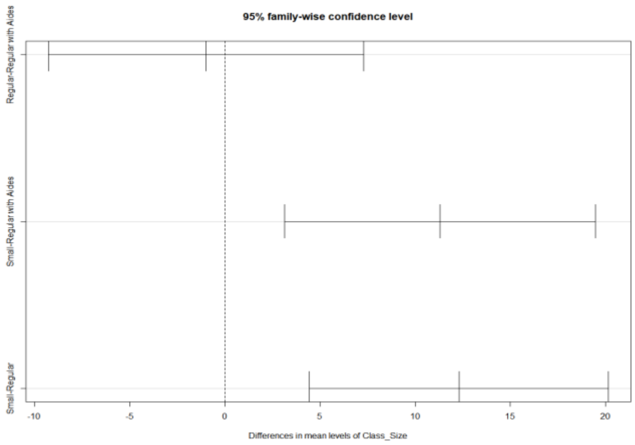

Figure 3: 95% family-wide confidence interval for class-type pairwise comparisons, Tukey's procedure

3.4.3 School Effects

The results of the F-test for school effects are shown in Table 3.

Table 3: Test for Factor Main Effects

Model	Degree of Freedom	SSE	F^*	P-value
Full	334	221371		
Reduced	335	230604	13.931	0.0002228

Since the p-value = 0.0002228, we reject H_0: $\beta_1 = \beta_2 = \ldots = \beta_{76} = 0$ at significance level 0.05 and conclude that there are school main effects.

4.0 Discussion

In this report, we presented our use of 2-way ANOVA to analyze the effect of class size on first-grade teachers' teaching performance in math in a stratified randomized experiment, using each school as a stratum.

4.1 Stratified Randomized Design

Exploratory analysis highlights the variability in teacher performance across distinct schools (Figure 1c). This variability is likely due to the similar demographic features within each school, paired with differing demographic features between them. Because of this high variability in median classroom performance between schools, blocking by school ID in our model helps to extract the precise effect of the class size treatment on the classroom median performance.

4.2 Exclusion of Interaction Terms

We explored the effect of including school-by-class size interactions in our model and concluded that interactions between the two factors would not add any value to our model. Philosophically, excluding interaction terms from the model fits the purpose of a stratified randomized experiment, because we are not primarily interested in the class size effect within individual schools, but rather its main effect across all schools. Eliminating interaction terms also results in fewer parameters to estimate and hence higher power of the test.

4.3 Class Size Effect

Model diagnostics suggested that our ANOVA assumptions were met. Results derived from fitting the model suggested significant difference in first-grade teacher's median math scores across different class sizes. Pairwise comparisons confirmed this result and further suggested that small classes outperformed both regular classes and classes with aides.

4.4 Causal Inference

The change of experimental units enables us to make causal statements regarding the effect of class size on teacher's performance in math education by satisfying the SUTVA and independence assumptions necessary for causal inferences:

SUTVA: Definition: *The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.*
The experimental unit used in the analysis satisfies the no-interference component of SUTVA – the assumption that the treatment applied to one unit does not affect the outcome for other units. On the basis of prior knowledge of school systems, it is realistic to assume that one teacher being assigned to a specific class size does not affect the teaching outcome of another teacher. The second component of SUTVA requires that individuals receiving a specific treatment cannot receive different forms of that treatment. In our case, due to the strict randomization implemented in the experiment, the class taught by one teacher is by nature homogenous with a class taught by another.

Independence Assumption:
Definition: *The assignment of treatment is independent of potential outcomes of experimental units.*
This assumption is met by using double randomization: One random assignment is that of teachers to classes. The second is of students to classes. The design ensures that high/low performance teachers or students were not systematically enriched in any class-size treatments. In light of this, systematic effects can be interpreted as the effects of class size.

Therefore, our analysis concludes that smaller class sizes has a positive average causal effect on a teacher's teaching outcome in math. This is different from the conclusion of Project I. SUTVA was not plausible when using individual students as experimental units. Interactions between students likely resulted in altered potential outcomes of one student due to the treatment assigned to another, thus violating SUTVA. In contrast, using teachers as experimental units does not rely on no-interference assumptions among students and makes our results credible evidence of causal class-size effects.

5.0 Reference

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models. New York: McGrawHill Education.
Suresh K. (2011). An overview of randomization techniques: An unbiased assessment of outcome in clinical research. Journal of human reproductive sciences, 4(1), 8–11. doi:10.4103/0974-1208.82352