Effects of Class Size on Math Performance in Tennessee

### Team ID: Team 6

#### NAME: Connor Rosenberg

#### NAME: Rongkui Han

#### NAME: Yuqing Yang

#### NAME: Nassim Ali-Chaouche

## 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 and conducted by the State Department of Education. Over 7,000 students from kindergarten to 3rd grade in 79 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 initiated as the students entered school in kindergarten and continued through third grade.

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 usually not to identify treatment effects within a single stratum, but rather treatment effects across all strata.

Because it is reasonable to expect systemic differences in educational outcome across schools, due to reasons such as demographics, school could be a confounding variable and influence the ooutcome 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 student, will be used as experimental units. The adjustment of experimental unit will facilitate the making of a causal statement as to the effect of class size on educational outcome, because the teacher experimental units more convincingly satisfy the stable unit treatment value assumption (SUTVA) and independence assumptions necessary for causal inferences. The detail of these assumtions will be dissected in the discussion section.

#### 1.2 Questions of Interest

1. Is there a significant difference in a first-grade teacher’s teaching performance in math across the three different class sizes?
2. Are teacher’s performances relatively stable between different schools? That is, does the school itself affect class average math scores?
3. Does our ANOVA model fit well with the data? In other words, are the analysis of variance assumptions satisfied?
4. Can we draw 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 truthfully reflects the class’ performance. In additon, 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:

* Completedness and balancedness of 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:

for , , and

Our model is constrained such that and . These two constraints imply the following impact on our parameter interpretations.

* represents the overall classroom median score across all treatment levels.
* represents the effect of each class size on the overall median math score.
* 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 which need estimation from 79 to a number that exceeds the sample size. This would impede us from fitting a powerful model. The second reason is that of relevance. The purpose of the analysis is to examine the main effects of class size on the median scaled math score achieved by a first-grade teacher’s class. We hope to derive conclusions applicable a general population. If we include an interaction term, our model now describes the effect of class sizes within individual schools. For the combination 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 sum of squares are no longer orthogonal. Therefore, we would use the general linear F-test for testing. The mechanism is to test whether specific components could be drop out of the full model. This is achieved by comparing sum of squared estimate of errors (SSE) under the full model with the SSE under the reduced models using the F statistic:

1. , where SSE(R) is SSE under the reduced model, is the degree of freedom for the reduced model, SSE(F) is SSE under the full model, and is the degree of freedom for the full model.
2. follows the F distribution, , under the null hypothesis ().
3. We would reject at level if , or if the p-value .

###### *2.2.3.1 Class Size Effects*

Then, we want to test whether or not class size effects are present:

:

: not all equal zero.

* Full model: .
* Reduced model is: .

If we reject at level , we conclude that the effects of class size are present.

###### *2.2.3.2 Pairwise Class Size Effects*

We will do pairwise comparisons among the three class sizes. The Tukey’s procedure will be used, and this procedure yields conservative result when sample sizes are unequal.

First, we define the differnce between two factor level means . The point estimate for is . Since and are independent, the variance of is . And the estimated variance of is .

Then, we do simultaneous testing:

If we control the family-wise confidence coefficient at level 1-, the confidence interval for is of the form:

where

We would check whether or not zero is contained in each interval. If zero is contained, we conclude ; otherwise, we conclude .

###### *2.2.3.3 School Effects*

Although the class size effects are of our primary interests, we also want to test whether or not school effects are present:

:

: not all equal zero.

* Full model: .
* Reduced model is: .

If we reject at level , we conclude the effects of school are present.

#### 2.3 Model Diagnostics

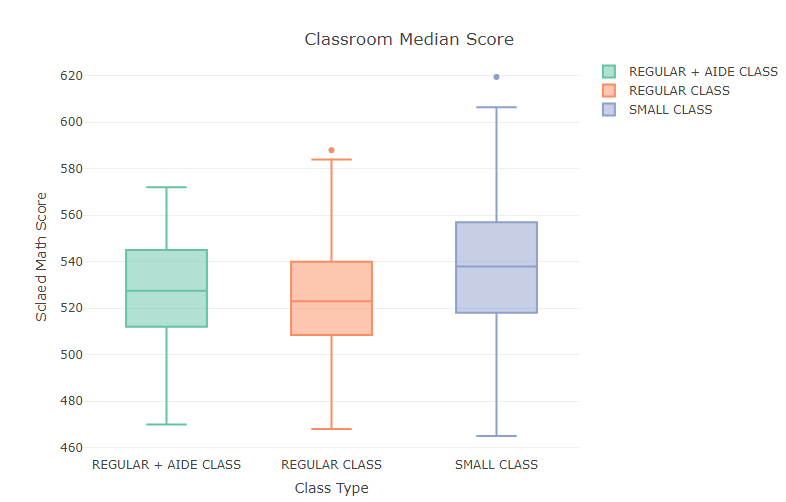
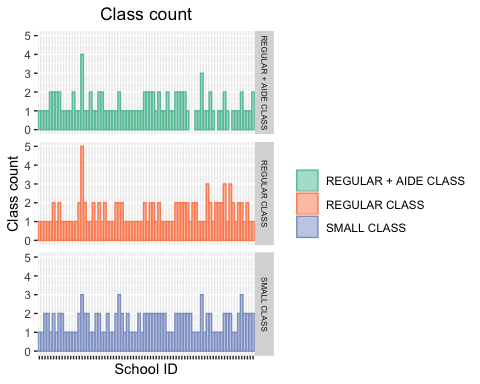
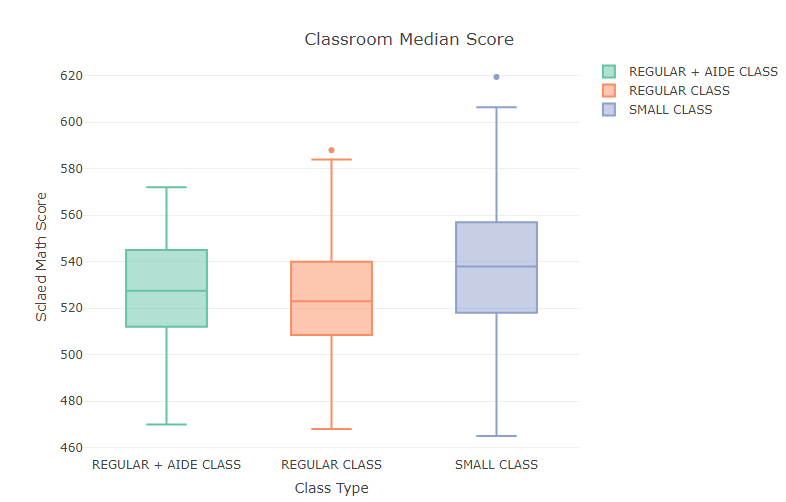
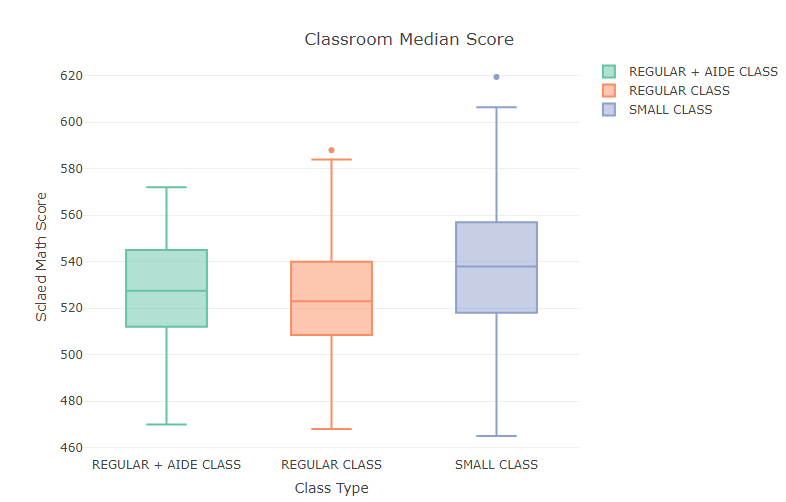
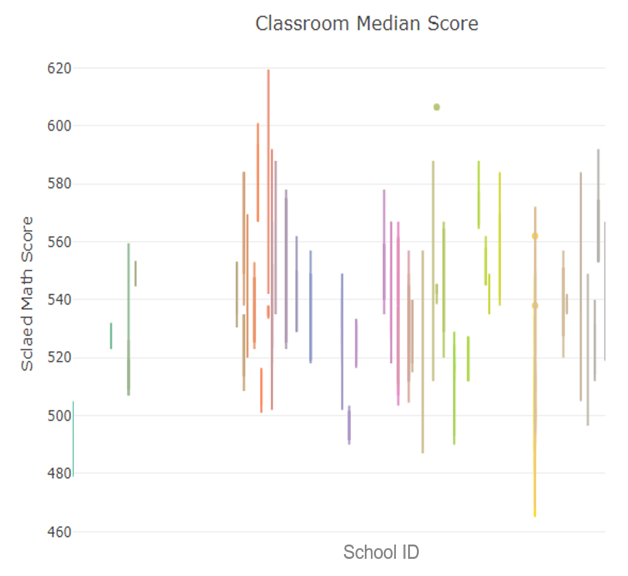
We will use Q-Q plot, histogram and the Shapiro-Wilk test inspect the normality of residuals. A scatter plot and a fitted-value-versus-residual scatter plot and the Levene test will be used to examine equality of residual variance. Independence of residuals and outlying data points will also be discussed.

## 3.0 Results

#### 3.1 Descriptive Analysis

After filtering out entries with missing datapoints in variables relevant to 1st grade performance, our dataset included median scaled 1st grade math scores from 338 teachers from 76 schools. Seventy-two out of 76 schools had least one complete set of the three different class sizes (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 treatments within them, we decided to retain 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, the boxplots of the classroom median scaled math scores for each school, highlights the high variability in teacher performance across distinct schools.

(a)

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

(b)

(c)

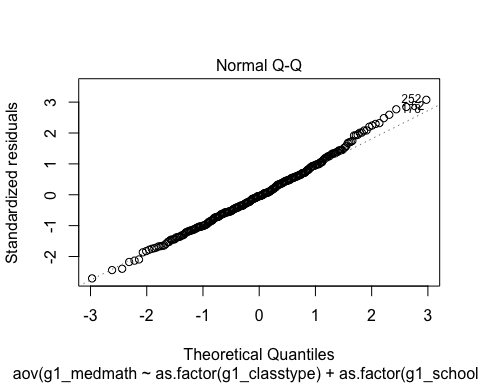
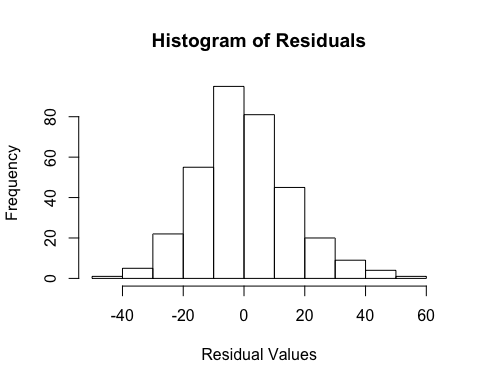
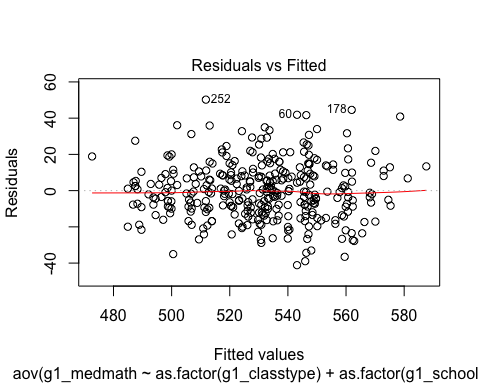
#### 3.2 Main Analysis

Table 1 describes the output for the two-way ANOVA model. The table includes the degrees of freedom, sum of squared error, mean squared error, and F test statistics and corresponding p-values. Both p-values for each variable, class type and school ID, fall below our specified 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 HERE.

Table 1: ANOVA Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variance Components | Degrees of Freedom | Sum of Squared | Mean of Squared |  | 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



(b)

(c)

(a)

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

##### 3.3.1 Normality:

From the Q-Q plot of the residuals (Figure 2a) , we can observe that most of that data points lie on a 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 points are only a few compared to the total amount of data. Thus, the normality assumption is largely satisfied from the Q-Q plot. A histogram is used to visualize the distribution of the residuals (Figure 2b). From the histogram we can observe that the distribution of the residuals looks symmetric about the mean and bell-shaped, similar to what is seen from a normal distribution.

To further test for normality of the errors, a Shapiro-Wilk test will be used on the distribution of the residuals. A Shapiro-Wilk test is used to test whether a distribution of data follows a normal distribution.

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

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

##### 3.3.2 Equal Variances:

The residuals are evenly spread out along the y-axis in a residual-versus-fitted value scatter plot (Figure 2c). Visually, the equal variance assumption is satisfied.

A Levene’s test is used to further test the equal variance assumption for both independent variables.

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

In the Levene’s test for both independent variables, both p-values (0.71 for class type and 0.92 for school ID) are greater than the significance level of 0.05. Thus for each variable, there is no evidence that residual variances are not equal across treatment groups. Thus, the equal variance assumption is satisfied.

##### 3.3.3 Independence:

This experiment is randomized in two ways and represents the best controlled environment to achieve independence of the results. First of all, teachers are randomly assigned to each class type: small, regular, and regular with aide. Second, every student is randomly assigned to each teacher. There is a possibility of the independence assumption not holding completely from the time of randomization to the time that the test scores were recorded. For instance, teachers may share materials, or parents of the same class may together decide to seek out tutoring for their children. However, given the randomized design of the experiment and the relatively large sample size, we may assume that the results are essentially independent.

##### 3.3.4 Outliers:

An outlier will be defined if it has a studentized residual value that is greater than 3 in absolute value. Entry 252, (teach ID 24475510, school ID 244755, small class size) was determined as an outlier (z = 3.13). Upon closer inspection of the entry, we did not find sufficient evidence to exclude it from the analysis.

#### 3.4 Hypothesis Testing

We use significance level 0.05 for all the following tests.

##### 3.4.1 Class Size Effects

The results of F-test for class type main effects is shown in the Table @ref(tab:class).

Test for Factor Main Effects

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Degree of Fredom | SSE |  | P-value |
| Full | 334 | 221371 |  |  |
| Reduced | 336 | 232391 | 8.3137 | 0.0002995 |

Since p-value = 0.0002995, we reject : at level of significance level 0.05. We conclude that there are class type main effects.

##### Pairwise comparisons of factor level means3.4.2 Pairwise Class Size Effects

Pairwise comparisons of factor level means

From Figure @ref(fig:tukey), we could see that all the confidence intervals do no contain zero. So we conclude that, at family-wise level , and , and , and are different. Small classes outperformed both regular classes and regular classes with aides.

Figure 2: 95% family-wide confidence interval for class-type pairwise comparisons, Tukey’s procedure

##### 3.4.3 School Effects

\*\*\*

The results of F-test for school effects are shown in Table @ref(tab:school).

Test for Factor Main Effects

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Degree of Fredom | SSE |  | P-value |
| Full | 334 | 221371 |  |  |
| Reduced | 335 | 230604 | 13.931 | 0.0002228 |

Since p-value = 0.0002228, we reject : at level of significance level 0.05. We conclude that there are school main effects.

## 4.0 Discussion

In this report, we presented our usage 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. This variability is likely due to the similar demographic features within schools, but various demographic features between them. For example, schools located in areas of high affluence may achieve better classroom performance since more students have access to academic support in addition to greater parent oversight. Similarly, schools who pull their students from less affluent areas may see worse classroom performance due to student food insecurity, lack of academic support, and other social deficiencies. 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 schoo-by-class size interactions in our model, and concluded that interactions between the two factors did not contribute significantly to the variance partitioning of the data. A model without interaction was used in all following analyses. Philosophically, excluding interaction terms from the model also 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 results in fewer parameters to estimate and hence higher power of the test.

#### 4.3 Class size Effect

Model diagnostics suggested that the dataset satisfied assumptions for ANOVA. Results derived from the fitting of the model suggested significant difference in a first-grade teacher’s median math scores across different class sizes. Pairwise comparisons suggested that small classes outperformed both regular classes and regular classes with aides. The model also revealed significant performance differences across schools.

#### 4.4 Causal Inference

This analysis enables us to make causal statements regarding the effect of class size on teacher’s performance in math education. This is made possible by using teachers as experimental units, thus 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 first 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 outcoome 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 in the experiment by using double randomization: One random assignment is that of teachers to classes. The second randomization is of students to classes/teachers. The design ensures that high/low performance teacher 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 size 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 outcome of one student due to the treatment assigned to another, thus violating SUTVA. In that case, rejections of the null hypothesis would not necessarily be convincing evidence of effects of class size; it may simply indicate the presence of peer effects. In contrast, using teachers as experimental units does not rely on no-interference assumptions among students. This makes the results reported here 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>