1 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). Classroom teachers were also randomly assigned to the classes they would teach. The interventions were initiated as the students entered school in kindergarten and continued through third grade.

Some of the key features of project STAR are:

1. *All Tennessee schools with K-3 classes were invited to participate.* Giving every school a chance to join the study helped ensure a diverse sample as well as rule out the possibility that class-size effects could be attributed to selection bias.
2. *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.* The within-school design provided built-in control for differences among schools in terms of resources, leadership, and facilities.
3. *Schools from inner-city, urban, suburban and rural locations were included in the experiment.* This feature guaranteed that samples would include children from various ethnic backgrounds and income levels.
4. *Students and teachers were randomly assigned to their class type.*
5. *Investigators followed the standard procedures for confidentiality in human subjects' research.*
6. *No children were to receive fewer services than normal because of the experiment.*
7. *Student achievement was to be tracked by standarized tests, which were carefully monitored.*

1.2 Questions of Interest

Our questions of interest are as follows:

1. Is there a significant difference in a first-grade teacher's average math scores 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 Analysis Plan

2.1 Population and study design

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 this study, each school can be viewed as a stratum. 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.

To expand on our previous findings, we will set teachers as the experimental unit, rather than the individual student. 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. The adjustment of experimental unit will enable us to make a causal statement as to the effect of class size on educational outcome.

2.2 Statistical Analysis

2.2.1 Descriptive Analysis

> Task 1: Explore math scaled scores in the 1st with teachers as the unit. Generate summary statistics (in forms of tables or plots) that you find informative, and explain them.

2.2.2 Main Analysis

2.2.2.1 Two-way Anova Model

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

$$Y_{ijk} = \mu_{..} + \tau_i + \beta_j + \epsilon_{ijk},\ i = 1, ..., a;\ j = 1, ..., b$$

Where:

- $a = 3,\ b = 76,$
- $\sum_{i=1}^{a} \tau_i = 0,\ \sum_{i=1}^{b} \beta_j = 0$
- Distribution assumotion: ϵ_{ijk} are independently and identically distributed as $N(0, \sigma^2)$.

and

- $\mu_{..}$ 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 sample size for the treatment consisting of the ith level for class size and the jth level of school will now be denoted by n_{ij}. And the total number of cases is $n_T = \sum_i \sum_j n_{ij}$. We estimate the population means by the corresponding sample means:

- $Y_{i..} = \hat{\mu}_{i.} = \frac{\sum_j \bar{Y_{ij.}}}{b}$, where $\mu_{i.} = \mu_{..} + \tau_i$
- $Y_{.j.} = \hat{\mu}_{.j} = \frac{\sum_i \bar{Y_{ij.}}}{a}$, where $\mu_{.j} = \mu_{..} + \beta_j$

Figure #, which shows the boxplots of the classroom median scores for each school, highlights the variability in teacher performance across each distinct school. This variability is likely due to the different 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, school's who pulls its population from less affluent areas may see worse classroom performance due to student food insecurity, lack of academic support, and other social deficiencies.

2.2.2.2 Model Diagnostics

2.2.2.3 Hypothesis Testing

Since we have unequal sample sizes, the factor effect component sum of squares are no longer orthogonal. Therefore, we would use the general linear F-test instead for the testing parts. The basic idea is to compare

SSE under the full model with SSE under the reduced model, and we want to test whether specific components could be drop out of the full model. The details are shown as below:

(1) Here the F-test statistic is: $F^* = \frac{\frac{SSE(R)-SSE(F)}{df_R-df_F}}{\frac{SSE(F)}{df_F}}$, 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.

(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 $< \alpha$.

2.2.2.3.1 Test for Interaction Effects

First, we want to test whether or not interaction effects are present. This would assess whether the effect of the factor class size differs across the stratum.

$H_0 : all\ (\tau\beta)_{ij} = 0$

H_a: not all $(\tau\beta)'_{ij}s$ equal zero

Here the full model is: $Y_{ijk} = \mu_{..} + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}$.

And the reduced model is: $Y_{ijk} = \mu_{..} + \tau_i + \beta_j + \epsilon_{ijk}$.

If we reject H_0 at level α, we conclude that there are interaction effects.

2.2.2.3.2 Test for Factor Main Effects

Class Size

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

H_0: $\tau_1 = \tau_2 = \tau_3 = 0$

H_a: not all $\tau'_i s$ equal zero.

(1) If there are interaction effects, then

- The full model is: $Y_{ijk} = \mu_{..} + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}$.
- And the reduced model is: $Y_{ijk} = \mu_{..} + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}$.

(2) If there are no interaction effects, then

- The full model is: $Y_{ijk} = \mu_{..} + \tau_i + \beta_j + \epsilon_{ijk}$.
- And the reduced model is: $Y_{ijk} = \mu_{..} + \beta_j + \epsilon_{ijk}$.

If we reject H_0 at level α, we conclude that the effects of class size are present.

School

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

H_0: $\beta_1 = \beta_2 = ... = \beta_{76} = 0$

H_a: not all $\beta'_i s$ equal zero.

(1) If there are interaction effects, then

- The full model is: $Y_{ijk} = \mu_{..} + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}$.
- And the reduced model is: $Y_{ijk} = \mu_{..} + \tau_i + (\tau\beta)_{ij} + \epsilon_{ijk}$.

(2) If there are no interaction effects, then

- The full model is: $Y_{ijk} = \mu_{..} + \tau_i + \beta_j + \epsilon_{ijk}$.
- And the reduced model is: $Y_{ijk} = \mu_{..} + \tau_i + \epsilon_{ijk}$.

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

Analysis of Class Size Effects

Because we are interested in the difference in the class size effects, we would do pairwise comparisons among the three class sizes. The Tukey's procedure will be used, and this procedure is conservative result when sample sizes are unequal.

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

Then, we do simultaneous testing:

$H_0 : D_{ii'} = 0$

$H_a : D_{ii'} \neq 0$

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

$\hat{D}_{ii'} \pm T \times s(\hat{D}_{ii'})$, where $T = \frac{1}{\sqrt{2}}q(1-\alpha; a, n_T - ab)$

We would check whether or not zero is contained in each interval. If zero is contained, we conclude H_0; otherwise, we conclude H_a.

3 Result

3.0.0.1 Hypothesis Testing

We use significance level 0.05 for all the following tests.

3.0.0.1.1 Test for Interaction Effects

The results of F-test for interaction effects is shown in Table 1.

Table 1: Test for Interaction Effects

Model	Degree of Fredom	SSE	F^*	P-value
Full	114	34612		
Reduced	260	81345	1.0543	0.3855

Since p-value = 0.3855, we can not reject $H_0 : all\ (\tau\beta)_{ij} = 0$ at level of significance level 0.05. We conclude that there is no interaction between these two factors.

As a result, we would revised the full model by excluding the interaction effects for the following tests. Also, we use this new full model for the main analysis.

3.0.0.1.2 Class Size

The results of F-test for class type main effects is shown in the Table 2.

Table 2: Test for Factor Main Effects

Model	Degree of Fredom	SSE	F^*	P-value
Full	334	221371		

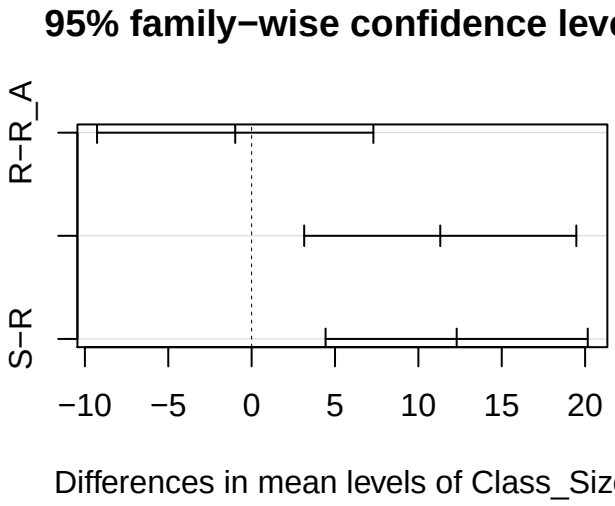

Figure 1: Pairwise comparisons of factor level means

Model	Degree of Fredom	SSE	F^*	P-value
Reduced	336	232391	8.3137	0.0002995

Since p-value = 0.0002995, we reject H_0: $\tau_1 = \tau_2 = \tau_3 = 0$ at level of significance level 0.05. We conclude that there are class type main effects.

3.0.0.1.3 School

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

Table 3: Test for Factor Main Effects

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

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

3.0.0.1.4 Analysis of Class Size Effects

From Figure 1, we could see that all the confidence intervals do no contain zero. So we conclude that, at family-wise level $\alpha = 0,05$, $\mu_{1.}$ and $\mu_{2.}$, $\mu_{2.}$ and $\mu_{3.}$, $\mu_{1.}$ and $\mu_{3.}$ are different. Moreover, the small classes outperformed both regular classes and regular classes with aides.

4 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. 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. 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 and regular classes with aides both outperformed regular classes without aides**. The model also revealed significant performance differences across schools, with the largest pairwaise difference being **XXX** in class median 1st grade scale math score.

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.