Tennessee Student/Teacher Achievement (STAR) Project

Replication Study for ECON 582

Harshvardhan, Id: 609162

University of Tennessee, Knoxville

Abstract

This replication paper is an attempt to recreate the results obtained by Krueger (1999). Through this exercise, I aim to rediscover the results obtained by the author, which would help cement my understanding. The paper studies 11,600 students in Tennessee public schools through kindergarten and K-3 levels. Project STAR was a large scale randomised trial costing $12 million over four years. Although the study was limited in some aspects, this was one of the closest attempts to study the impact of class size on teaching and learning outcomes. Krueger (1999) found that students in smaller classes perform better than students in larger classes. Further, students with teaching aides didn't perform significantly better than those without teaching aides. I will reproduce their tables and figures, and elaborate on inferences.

Keywords: Tennessee STAR Project, Replication, Ordinary Least Squares, Two-stage Least Squares

1 Introduction

The Tennessee Student/Teacher Achievement (STAR) Project was a large scale randomised experiment on class size conducted in Tennessee schools. The students and teachers were randomly assigned to one of the three class types: small class (13-17 students per teacher), regular-size classes (22-25 students) and regular classes with teacher's aide (22-25 students). Over the four years, 11,600 students were part of the study from 80 schools. The randomisation (random assignment of teachers and students to one of these classrooms) happened at the school level.

This paper and STAR project aimed to discover the importance and impact of class size on learning outcomes. Colloquially speaking, it is hypothesised that smaller classes, i.e. classes with a low student-teacher ratio, have better learning outcomes. Krueger (1999) aimed to quantify the learning outcomes based on the class size using the Project

STAR dataset. Stanford Achievement Test (SAT) and Tennessee Basic Skill First (BSF) tests were used as the proxy variables for student achievement or learning outcome.

Krueger (1999) considers that class-size dummy variable and regular-size-with-aide dummy variables together can explain the effect on student achievement — while controlling for other student-teacher related attributes (controlled covariates) and the school where the student is enrolled. The average percentile score with SAT measures the outcome of student achievement. Class size dummy and regular-class-with-aide are taken directly from the project database. The control variables are included as covariates such as gender, age, among others. School-related effects are included as a separate control variable in the study.

This results in the following regression model:

$$Y_{ics} = \beta_0 + \beta_1 * SMALL_{cs} + \beta_2 REG - A_{cs} + \beta_3 X_{ics} + \alpha_s + \epsilon_{ics}, \tag{1}$$

where

- Y_{ics} is the average percentile score on the SAT test of student i in class c at school s,
- $SMALL_{cs}$ is a dummy variable indicating whether the student was assigned to a small class that year,
- $REG - A_{cs}$ is a dummy variable indicating whether the student was assigned to a regular-size class with an aide that year,
- X_{ics} is a vector of student and teacher covariates used as control,
- α_s is the school effect,
- ϵ_{ics} is the error term.

In my notation, Y_{ics} is $\mathbf{Y}$ and all the other variables are part of $\mathbf{X_1}$.

Thankfully this was a randomised experiment. If this were not a randomised experiment, we would need several assumptions to infer the effects reliably. The first assumption would be about the conditional independence of each variable included in the study. This implies that we would assume that there is no correlation between the variables included in the study, i.e. $Cov(X_i, X_j) = 0$ for all $i \neq j$. We would also need to assume that errors are independent of the variables included in the study.

There could be multiple omitted variables in the study. There was no measure of students' inherent ability in the study. Some students would likely be more intelligent than others, and there is no way to know or measure that in this regression model. A proxy like IQ scores could've been included. This inherent intelligence will affect student outcomes and affect which school the student joins as "smart" students tend

to be clustered in some schools more than the others. Further, there have been many studies indicating or at least hypothesising that gender, age, and race can affect inherent intelligence.

Randomisation would not solve this issue altogether. The school effect is captured in one of the variables, but the individual traits (like gender, age and race) have not been randomised (or at least the study doesn't mention that). Therefore, the estimates are likely biased.

Rest of this paper is organised as follows. Section 2 discusses some details on data and how I managed the project. Section 3 showcases all the reproduced tables and figures from Krueger (1999) study. Section 4 has the exploratory discussions on some tables and what can we conclude from each of them. Section 5 has some notes of regression experimental design and its conclusions. Section 6 is on limitations of Krueger (1999) study. Finally, some concluding remarks are added in section 7.

2 Data and Codebase Organisation

Complete dataset was available at Harvard Dataverse (Achilles et al., 2008). The databse contained raw student and school level data from the longitudinal experiment. Primary student-level data is available for 11,601 students who participated in the study for at least one year. Demographic variables, school and class identifiers, school and teacher information, experimental conditions, achievement test scores and motivation and self-concept scores are available. For more details, see Achilles et al. (2008).

For this study, I designed a working directory with the following folders:

- **DO files:** for all the coding files (such as STATA's `.do` files),
- **DTA files:** for the input dataset (STAR Project Dataset),
- **Figures:** for all the results to be stored,
- **TEX:** for all the LaTeXfiles,
- **others:** for any other file that doesn't belong to any of the above category.

I wrote modular codes such that each code file `.do` does only one analyses, i.e. either generates a table or a figure. For version control and backup, I used GitHub. For extra protection, I stored all the analyses files locally on a private Dropbox folder as well. The Github repository for this project is located at `https://github.com/harshvardhaniimi/krueger1991-replication`. All the codes to generate the figures and tables are present there.

3 Replicated Tables and Figures From Krueger (1999)

This section contains tables replicated from Krueger (1999). All the coming sections will refer to these tables for insights.

3.1 Table I

	Small	Regular	Regular + Aide	Joint P-value
Free Lunch	0.47	0.48	0.50	0.09
White/Asian	0.68	0.67	0.66	0.26
Age in 1985	5.44	5.43	5.43	0.33
Attrition Rate	0.49	0.52	0.53	0.02
Class Size	15.12	22.38	22.77	0.00
SAT Percentile Score	54.73	49.95	49.99	0.00

Tab. 1: Comparison of mean characteristics of treatments and controls for the students who entered STAR programme in kindergarten.

	Small	Regular	Regular + Aide	Joint P-value
Free Lunch	0.59	0.62	0.61	0.52
White/Asian	0.62	0.56	0.64	0.00
Age in 1985	5.78	5.86	5.88	0.03
Attrition Rate	0.53	0.51	0.47	0.07
Class Size	15.87	22.71	23.46	0.00
SAT Percentile Score	49.52	42.90	48.02	0.00

Tab. 2: Comparison of mean characteristics of treatments and controls for the students who entered STAR programme in Grade 1.

	Small	Regular	Regular + Aide	Joint P-value
Free Lunch	0.66	0.63	0.66	0.60
White/Asian	0.53	0.54	0.44	0.00
Age in 1985	5.88	5.91	5.94	0.41
Attrition Rate	0.37	0.34	0.35	0.58
Class Size	15.50	23.71	23.59	0.00
SAT Percentile Score	46.56	45.45	41.84	0.01

Tab. 3: Comparison of mean characteristics of treatments and controls for the students who entered STAR programme in Grade 2. Joint P-value for Age in 1985 doesn't exactly match, which is likely due to replication issues.

	Small	Regular	Regular + Aide	Joint P-value
Free Lunch	0.60	0.64	0.69	0.04
White/Asian	0.66	0.57	0.55	0.00
Age in 1985	5.95	5.93	5.99	0.50
Class Size	15.97	24.05	24.43	0.01
SAT Percentile Score	47.86	44.51	41.54	0.01

Tab. 4: Comparison of mean characteristics of treatments and controls for the students who entered STAR programme in Grade 3. Joint P-value for Age in 1985 doesn't exactly match, which is likely due to replication issues.

3.2 Table II

Variable	Grade Entered STAR Programme			
	K	1	2	3
Free Lunch	0.46	0.29	0.58	0.18
White/Asian	0.66	0.28	0.18	0.27
Age in 1985	0.44	0.12	0.43	0.48
Attrition Rate	0.01	0.37	0.85	NA
Actual Class Size	0.00	0.00	0.00	0.00
SAT Percentile Score	0.00	0.00	0.47	0.00

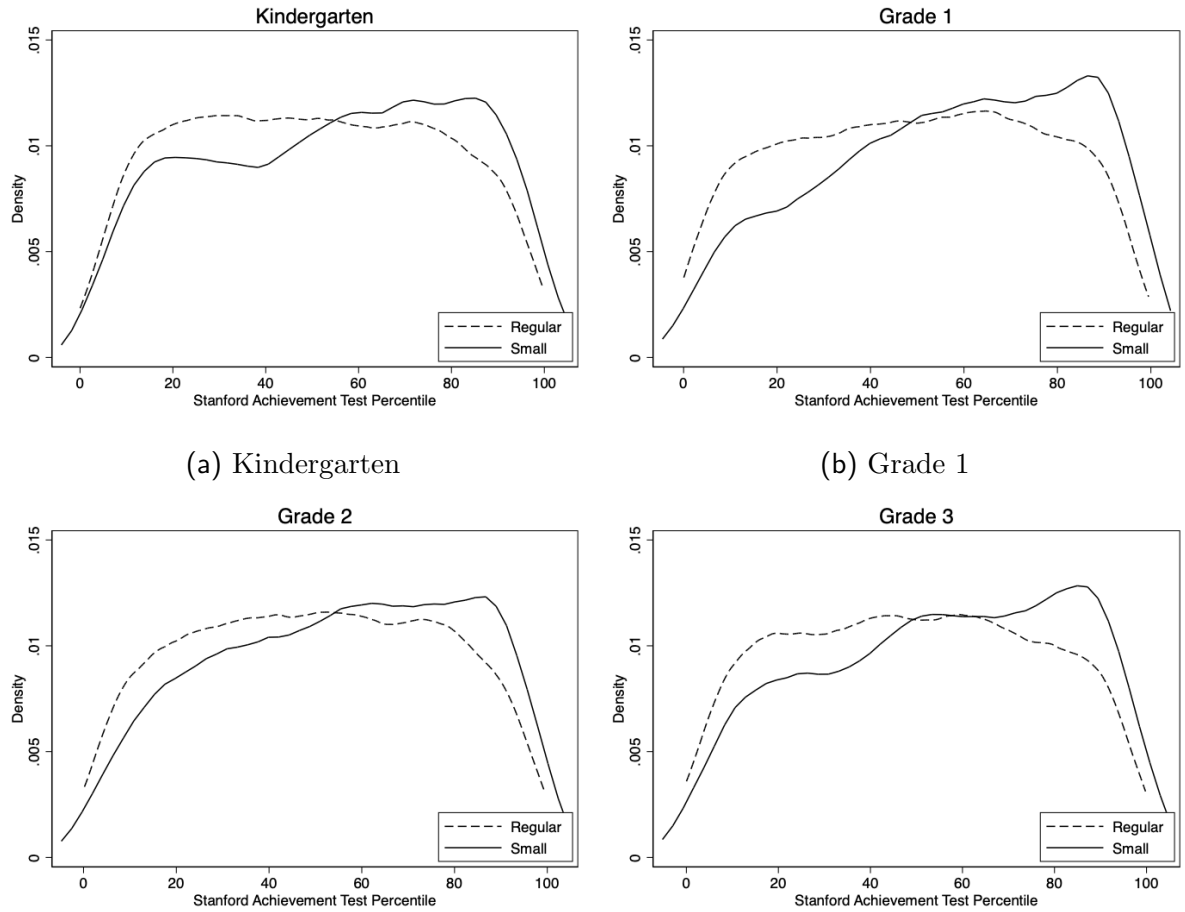
Tab. 5: P-values for tests of within-school differences between small, regular and regular classes with aide. Some values do not match exactly (notably age) which is likely due to replication errors.

3.3 Table III

	Assignment Group in First Grade		
Actual class size in first grade	Small	Regular	Regular with Aide
12	24	0	0
13	182	0	0
14	252	0	0
15	465	0	0
16	256	16	0
17	561	17	0
18	108	36	0
19	57	76	57
20	20	200	120
21	0	378	378
22	0	594	330
23	0	437	460
24	0	384	264
25	0	175	225
26	0	130	234
27	0	54	108
28	0	28	56
29	0	29	58
30	0	30	30
Average	15.7	22.7	23.4

Tab. 6: Distribution of children cross actual class sizes in grade 1, assigned randomly. Replicated from Table III.

3.4 Figure I



(a) Kindergarten (b) Grade 1

(c) Grade 2 (d) Grade 3

Fig. 1: Density plot of SAT percentile distribution for each class by size and grade. Regular class with aide are considered same as regular classes.

3.5 Table V Regenerated

	OLS: Actual Class Size				Reduced Form: Initial Class Size			
Explanatory Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Small class	4.72*	5.34***	5.32***	5.34***	4.72*	5.34***	5.32***	5.34***
	(2.20)	(1.26)	(1.22)	(1.19)	(2.20)	(1.26)	(1.22)	(1.19)
Regular/Aide class	-0.03	0.22	0.44	0.26	-0.03	0.22	0.44	0.26
	(2.24)	(1.12)	(1.09)	(1.06)	(2.24)	(1.12)	(1.09)	(1.06)
White/Asian			8.31***	8.41***			8.31***	8.41***
			(1.35)	(1.36)			(1.35)	(1.36)
Gender (Girl)			4.49***	4.41***			4.49***	4.41***
			(0.63)	(0.63)			(0.63)	(0.63)
Free Lunch			-13.16***	-13.08***			-13.16***	-13.08***
			(0.78)	(0.77)			(0.78)	(0.77)
White Teacher				-1.22				-1.22
				(2.15)				(2.15)
Teacher's experience				0.26*				0.26*
				(0.10)				(0.10)
Master's degree				-0.49				-0.49
				(1.08)				(1.08)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	0.01	0.25	0.31	0.31	0.01	0.25	0.31	0.31

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Tab. 7: OLS and reduced form estimates of the effect class-size assignment on average percentile on SAT for Kindergarten students.

	OLS: Actual Class Size				Reduced Form: Initial Class Size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Small Class	8.60***	8.45***	7.93***	7.61***	7.59***	7.21***	6.83***	6.60***
	(1.98)	(1.21)	(1.17)	(1.17)	(1.77)	(1.14)	(1.10)	(1.10)
Regular/aide Class	3.42	2.20*	2.21*	1.77	1.94	1.71*	1.66*	1.53*
	(2.05)	(0.99)	(0.97)	(0.97)	(1.12)	(0.80)	(0.76)	(0.76)
White/Asian			6.99***	6.98***			6.87***	6.86***
			(1.18)	(1.19)			(1.18)	(1.19)
Gender (Girl)			3.79***	3.83***			3.76***	3.80***
			(0.56)	(0.56)			(0.56)	(0.56)
Free Lunch			-13.43***	-13.53***			-13.59***	-13.70***
			(0.87)	(0.87)			(0.88)	(0.88)
White Teacher				-4.05*				-4.14*
				(1.95)				(1.97)
Teacher experience				0.06				0.07
				(0.06)				(0.06)
Master's Degree				0.34				0.48
				(1.07)				(1.10)
School Fixed Effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	0.02	0.24	0.30	0.30	0.01	0.23	0.29	0.30

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Tab. 8: OLS and reduced form estimates of the effect class-size assignment on average percentile on SAT for Grade 1 students.

	OLS: Actual Class Size				Reduced Form: Initial Class Size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Small Class	5.96**	6.28***	5.83***	5.75***	5.30**	5.46***	5.27***	5.24***
	(1.98)	(1.29)	(1.23)	(1.22)	(1.70)	(1.16)	(1.10)	(1.09)
Regular/aide Class	2.07	1.97	1.74	1.67	0.59	1.56	1.29	1.30
	(2.07)	(1.10)	(1.07)	(1.06)	(1.25)	(0.87)	(0.82)	(0.81)
White/Asian			7.05***	7.06***			6.98***	7.00***
			(1.18)	(1.18)			(1.19)	(1.19)
Gender (Girl)			3.30***	3.27***			3.30***	3.27***
			(0.60)	(0.60)			(0.60)	(0.60)
Free Lunch			-13.55***	-13.55***			-13.68***	-13.67***
			(0.72)	(0.72)			(0.73)	(0.73)
White Teacher				0.43				0.46
				(1.75)				(1.77)
Teaching Experience				0.10				0.10
				(0.06)				(0.07)
Master's Degree?				-1.00				-1.10
				(1.06)				(1.05)
School Fixed Effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	0.01	0.22	0.28	0.28	0.01	0.21	0.28	0.28

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Tab. 9: OLS and reduced form estimates of the effect class-size assignment on average percentile on SAT for Grade 2 students.

	OLS Actual Class Size				Reduced Form: Initial Class Size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Small Class	5.32**	5.57***	5.02***	5.12***	5.50***	5.42***	5.31***	5.39***
	(1.93)	(1.22)	(1.20)	(1.22)	(1.47)	(1.08)	(1.03)	(1.06)
Regular/aide Class	-0.28	-0.19	-0.36	-0.48	-0.38	0.09	0.09	0.06
	(1.96)	(1.13)	(1.12)	(1.10)	(1.18)	(0.86)	(0.81)	(0.80)
White/Asian			6.10***	6.09***			5.95***	5.95***
			(1.45)	(1.44)			(1.44)	(1.43)
Gender (Girl)			4.14***	4.14***			4.15***	4.15***
			(0.66)	(0.66)			(0.66)	(0.66)
Free Lunch			-13.03***	-13.00***			-13.21***	-13.20***
			(0.81)	(0.81)			(0.82)	(0.82)
White Teacher				0.37				-0.05
				(1.80)				(1.80)
Teacher's Experience				0.06				0.05
				(0.06)				(0.06)
Master's Degree				0.74				0.56
				(1.18)				(1.18)
School Fixed Effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	0.01	0.17	0.22	0.22	0.01	0.16	0.22	0.22

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Tab. 10: OLS and reduced form estimates of the effect class-size assignment on average percentile on SAT for Grade 3 students.

3.6 Table VII Regenerated

Grade	OLS	2SLS	Sample Size
K	-0.62	-0.72	5840
	(0.14)	(0.14)	
1	-0.85	-0.67	6455
	(0.13)	(0.15)	
2	-0.60	-0.53	6011
	(0.12)	(0.13)	
3	-0.61	-0.67	6124
	(0.13)	(0.13)	

Tab. 11: OLS and 2SLS estimates of the effect of class size on achievement on dependent variable of average percentile score on SAT for different grades. The models control for school fixed effects, students' race, gender and free lunch status, teachers' race, experience and education. In parentheses are robust standard errors that allow for correlated errors among the students.

4 Exploratory Investigations

4.1 How Random was Initial Assignment?

As discussed before, this experiment does not have a control group. In a perfect world with a perfect experimental set up, we would look at the same student being assigned to different treatments, i.e. assigned to a small class, a regular class or a regular class with aide. However, if the randomisation was plausible then we could expect that the students belonging to different groups didn't show any major differences with each other *on average*. Table 1 to 4 examine this.

Students were assigned to different groups when they joined the program. If we compare their SAT percentile scores and other controls, we can conclude if the students in different groups did have any difference with each other. As seen in Tables 1 to 4, these differences are significant in certain groups. Therefore, it is essential that we control for these in our regression model.

Schools also have a strong impact on different outcomes and memberships. For example, a school in economically impoverished neighbourhood would have more students sign up for free lunch. Therefore, we should inspect the difference between the groups conditional on school fixed effects. Table 5 presents the results of joint F-tests of difference between small, regular and regular classes with aide for different variables and students' joining grade.

None of the background variables (free lunch, being white or asian, and age) are sig-

nificantly different between the treatment assignment at ten per cent level. This evidence suggests that the assignment was fairly random. Krueger (1999) also ran the same analysis combining all the four grades together. All the background variables do not show any significant signs of association (Krueger, 1999).

4.2 How Distributed were Assignment Groups?

Not all classes in the same group had equal number of students. As evident from Table 6, small classes had fewer students than regular classes. Indeed, the minimum and maximum class size for small treatment group was 12 and 20 respectively. Some assigned regular classes had fewer than 20 students but a majority of regular classes had 20 or more students.

This provides evidence to the fact that the classes were assigned as they were defined by the researchers. This is another critical check of assumptions mandatory for causal inferences from this experimental data.

4.3 Are Smaller Classes Better?

Figure 1 shows the kernel density of the average test scores for different class types and grades. Each kernel density plot compares the SAT percentile scores between small and regular class. Regular class includes regular classes with aide in these plots.

As see from the figures, the solid line overtakes dashed line around SAT percentile score of 50. Students in smaller classes perform better than students in regular classes. In fact, most students performing better than median actually belong to small classes and vice versa.

Krueger (1999) tests robustness of this conclusion using more advanced tools. I will discuss those tests and regression models in following subsections.

5 Regression: Design and Results

Krueger (1999) uses the following regression model to estimate the effect of schools' resources on student achievement:

$$Y_{ics} = \beta_0 + \beta_1 SMALL_{cs} + \beta_2 REG/A_{cs} + \beta_3 X_{ics} + \alpha_s + \varepsilon_{ics}, \tag{2}$$

where Y_{ics} is the average SAT percentile score of student i in class c at school s; $SMALL_{cs}$ (REG/A_{cs}) is an indicator variable with value one when the student is in small class (regular class with aide) c at school s; X_{ics} is set of control variables for student's background and school characteristics. Because many characteristics depend on school, α_s captures their effect.

5.1 Design

Ordinary Least Squares (OLS) The above regression equation can be estimated using ordinary least squares. Because some assignments changed over the years, Krueger (1999) estimates the regression models again using the student's initial assignment. Those that include initial assignment are labelled "reduced form" models as the initial assignments are actually an excluded variable in the regression, correlated with actual class size.

To interpret the causal inference, we need to make certain assumptions about our experiment design. While we accept that this study is very close to random experiment, we still might have issues. For example, the stochastic error ε_{ics} would include variables omitted in regression design like teacher's influence (some teachers are better at pep-talks than others). We will have to assume partial conditional independence.

To interpret the coefficient of small classes and teacher aides as causal effects, we will need to assume that belonging in a small class is independent to such omitted variables.

Two-stage Least Squares (2SLS) As described in Table 6, different class groups had different number of students. In fact, there was an overlap between the groups too. 2SLS models can take this variability between the groups into account by separately estimating two equations.

$$CS_{ics} = \pi_0 + \pi_1 S_{ios} + \pi_2 R_{ios} + \pi_3 X_{ics} + \tau_{ics} \tag{3}$$

$$Y_{ics} = \beta_0 + \beta_1 CS_{ics} + \beta_2 X_{ics} + \alpha_s + \varepsilon_{ics} \tag{4}$$

CS_{ics} is the actual number of students in the class, S_{ios} is an indicator variable for small class assignment (initial) and R_{ios} is another indicator variable for regular class assignment (initial). With this setup, the variations are captured by the initial assignment. Because this assignment was random, this *excluded* instrument would not be correlated with ε_{ics}. However, as there were non-random transfers from one group to another over the years, this assumption is controversial (except for kindergarten, where there were no switches).

Therefore, the regression estimates match exactly for 2SLS and OLS models at the kindergarten level.

5.2 Results

Regression results from OLS are presented in Table 7 to 10. It can be concluded that students in smaller classrooms perform better than students in larger classrooms. Students in small classroom score 5 percentile points higher in kindergarten, 8.6 percentile points higher in grade 1 and between 5 and 6 percentile points in grade 2.

For kindergarten, there is no difference between between the reduced form and OLS estimates as students couldn't change the assignment before the academic year. As seen

in column 4 (which controls for background variables), being in a small class increases SAT percentile score by 5.34 points, and the increase is significant with 99.9% confidence. Having a teaching aide in class doesn't increase achievement score.

In grade 1 students, the increment is 7.61 percentile points (based on current assignments) and 6.60 percentile points. Having a teaching aide increases score by 1.5—1.7 but the increment is not statistically significant. In grade 2, students in smaller class perform 5.24—5.75 points better and this difference is statistically significant. Again, having a teaching aide doesn't bolster performance.

In grade 3 students, the story is repeated. Small class students perform 5.12—5.39 points better which is statistically significant and there is no major gain when using teaching aide.

The 2SLS estimates (presented in Table 11) show that 2SLS estimates are slightly larger. (Also concluded from results in Krueger (1999). In my replication, the coefficients are slightly smaller for grades 1 and 2.) These results indicate that those who attend smaller classes tend to score higher at the end of first year when they enter the program.

6 Limitations

A major limitation to my mind of this study is that there is no measurement of teaching quality of the teachers. It is possible and likely that different teachers had different levels of teaching motivation and some could explain concepts better than others. The study didn't capture any information on teaching quality achievement. If a teachers' race, experience and education are orthogonal to their teaching quality (which is plausible), there is no way to know. For more experiment design limitations, see Hanushek (1999).

If the difference is this significant, I would have explored the regression discontinuity design as well. If all other factors are controlled for, we could directly estimate the difference between being in a small vs regular class.

Furthermore, Nye et al. (1999) question the validity of results due to high attrition rates in the experiment. More than half of the students who joined in kindergarten had left the experiment by grade 3. However, they make a strong case that such inconsistencies were not significant and the experiment's results were still valid.

7 Concluding Remarks

In this short replication paper, I attempted to rediscover results produced by Krueger (1999). I found that most tables and figures could be replicated using the methods as described in the paper. In a sentence, we can conclude that *students in small classes perform better than regular classes* and *having teacher's aide in class does not have any significant effect on student achievement.* Finally, I discussed the limitations of this study and how it could be improved.

Acknowledgement

Sincere thanks to Dr Carruthers and Cathy Wu for their kind support and help through the project. Most of this code base was written by Dr Carruthers, Adrienne Sudbury, Ge Wu and others. I have recycled parts of it and added more for the purpose of this project. The complete code base is available at `https://github.com/harshvardhaniimi/krueger1991-replication`.

References

Achilles, C., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., Johnston, J., and Word, E. (2008). Tennessee's Student Teacher Achievement Ratio (STAR) project.

Hanushek, E. A. (1999). Some findings from an independent investigation of the tennessee star experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis*, 21(2):143–163.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, 114(2):497–532.

Nye, B., Hedges, L. V., and Konstantopoulos, S. (1999). The long-term effects of small classes: A five-year follow-up of the tennessee class size experiment. *Educational Evaluation and Policy Analysis*, 21(2):127–142.