Undergraduate Economic Review

Volume 15 | Issue 1 Article 4

2018

The Effects of Student-Teacher Ratio on Test Scores: Applying Ceteris Paribus to California and Massachusetts Schools

Laira Aggarwal
Claremont McKenna College, laggarwal19@cmc.edu

The Effects of Student-Teacher Ratio on Test Scores: Applying Ceteris Paribus to California and Massachusetts Schools

Abstract

This paper seeks to analyze the impact of student-teacher ratio on test scores in California and Massachusetts. Since student-teacher ratio is just one of the variables affecting students' learning outcomes, other attributes were taken into account for a comprehensive analysis. These attributes included percent of English learners, average district income, percent of students on free or reduced lunch, and expenditures per student. The data sets for both states were assessed both inherently and with ceteris paribus approach. The results indicated that while student-teacher ratio does affect test scores, other classroom variables have a significantly greater influence on students' learning outcomes.

Introduction

The United States, despite its popularity and reputation as a destination for most international students, has been suffering from its faltering quality of the secondary education. As an effort to alleviate the declining quality of education, numerous districts took an approach of lowering class sizes. Most notably, the Student/Teacher Achievement Ratio (STAR) Experiment in Tennessee exhibited that students in smaller classes, from Kindergarten to third grade level, had higher achievement than those in larger classes. Furthermore, based on the results of Swedish policy reforms, Fredriksson and Ockert also found that the student performance increased by 2.6 percentile ranks. While many argue that the reduction in class sizes led to a greater scholastic achievement, others, such as Hanushek (1999), have criticized that there seems to be a minor "systematic gain from general reductions in class size" and that the effect of such programs will depend more on the quality of teachers than on the class size reduction. Hence, Mitchell and Mitchell (1999) concluded from their study that California's Class Size Reduction (CSR) program, controlling demographic variables, had a slight, insignificant improvement in students' achievement. Considering that lowering the student-teacher ratio (STR) can be fiscally burdensome, we attempt to assess the effectiveness of the class reduction programs in improving students' outcomes, holding other variables constant ("Ceteris Paribus").

The purpose of this paper is to evaluate whether reducing the class size significantly improves the students' test scores. We proceed as follows: Section II delves into the methods of our analysis, specifically on the selection of our data sets, comparison of the data, and causal effect and selection bias. Section III exhibits the results and expounds on our findings. Finally, Section IV summarizes the results from our research and establishes the correlation between test scores and the STR.

Methods of Analysis

A. Selection of data sets and relevant attributes

The causal impact this paper seeks to examine is between class size and student learning outcomes. Two samples were examined: one from California (420 schools) and one from Massachusetts (220 schools). To maintain consistency in the studies of the two states, standardized test scores of grades 5 and 4 students were taken for California and Massachusetts respectively. Class size was measured using the average student-teacher ratio (STR) attribute, and the data for both the states was sorted into small class (< 20 students per teacher) and large class (>20 students per teacher). For this report, the explanatory variable was class size. This is because, the purpose of the experiment was to check whether class sizes affect test scores. The treatment group was small class sizes and the control group was large class sizes.

Additionally, other relevant attributes apart from student –teacher ratio were taken into consideration to check for their effect on test scores. These include percent of English learners (EL_Pct), average income in the school district (AVG_INC), percent qualifying for reduced-price lunch (MEAL_PCT), and expenditures per student (EXPN_STU). The percentage of students still learning English in classes tends to affect their performances in tests (which are in English language). Moreover, average income and meal plan information (students on reduced and free lunch plans) provides information about poverty, and poverty does indeed affect the test scores of students. Finally, the attribute of expenditures per student was selected because the resources available to students can affect their performance.

B. Using t-distribution and confidence intervals

The main method of analysis in this report comes from using Student's t-distribution and constructing confidence intervals. t-distribution was used because population variance was unknown. It is crucial to use the t-table to find the number of standard deviations from the mean and use that value accordingly to calculate the respective confidence intervals. For sample size greater than 100, it is safe to use the z-table because in large sample sizes t-distribution converges to normal distribution and thus values from that table can be approximated. In many of the cases that were analyzed, there were usually more than 100 observations, and so z-table could be used. However, many times in the Massachusetts data, smaller samples were provided, and so t-distribution became persistent. Additionally, it is known that if two samples have different variances then the following equation must be used to find the degrees of freedom, V:

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1+1} + \frac{\left(S_2^2/n_2\right)^2}{n_2+1}} - 2$$

Where s is the variance and n is the number of observations in the sample. It was necessary to do this for every single difference in means as it is highly unlikely that two samples with have the same variance. However, many times the degrees of freedom allowed the test to converge to the normal distribution.

C. ***Variables***

The treatment and control variables in this study will be referred to as Y_1 and Y_0, respectively. A school either has a class size of greater than 20, which is considered the control Y_0, or it has a class size of less than 20, which is considered the treatment Y_1. In a perfect world where data about a single district, *i*, is known, the treatment effect would be found by $Y_{1i} – Y_{0i}$. However, since the data set does not have this information, comparisons have to be made between schools. In this case, the treatment effect is found by finding the difference in test score between district *i*, and district *j*. The treatment effect will now be described as $Y_{1i} – Y_{0j}$, assuming *i* is in the treatment group and *j* is in the control group.

Equally important to be aware of is the effect of selection bias. Part of the reality of comparing effects of treatments between districts is that instead of the effect of no treatment, Y_0, staying constant, it changes with every district. In Appendix A, Y_0 is the potential test score with a large class, or no treatment. The difference between Y_{0i} and Y_{0j} for any two district is called the selection bias. The selection bias can affect the results depending on the effect of omitted variables. It is important to acknowledge that neither $Y_{0i}– Y_{0j}$ nor $Y_{1i} – Y_{0i}$ is actually visible or able to be directly measured given the data. Angrist and Pischke point this out in Table 1.2, which has been recreated in the context of this study using the data that was available. The two districts were chosen randomly using a random number generator. For any two districts being compared, only about half of the table's values will be known, since only one observation, $Y\text{-}_{1i}$ or Y_{0i} , was made per district. This is important to remember moving forward, as these limitations in data will need to be addressed in order to clarify the result.

D. Different classroom sizes comparison and correlation of relevant attributes

The average test scores of students in small class sizes and large class sizes were calculated for both California and Massachusetts using summary statistics in Excel. The difference in means of test scores for the two class sizes was computed for both the states, to understand the impact of student-teacher ratio on test scores. This was done taking into account the summary of mean of class sizes and test scores that allowed for a test of significance between the two variables. If the confidence interval (95%) for difference in means did not contain the null (0 in this case), then the difference in test scores between small and large classes in California and Massachusetts was statistically significant, thus implying that student teacher ratio affects test scores. However, it's crucial to note that this analysis does not take into account other relevant attributes such as EL_Pct, AVG_INC, MEAL_PCT, and EXPN_STU that might possibly be somewhat responsible for the variability in test scores.

Additionally, a 6 X 6 correlation matrix for each sample that included test scores, student- teacher ratio, and other four attributes was created for both the states. A correlation matrix provides information about proximity of the relationship between each possible combination of imputed variables. It elucidates if one variable has a stronger relationship with test scores than the others.

If the results show that indeed other variables are statistically significant, and they show strong correlation with test scores then the concept of "ceteris paribus" (to hold everything else constant) addressed by Angrist and Pischke cannot be applied here, because other variables cannot be held constant if they in fact are influencing students' learning outcomes.

E. "Ceteris Paribus" approach

Due to the lack of ceteris paribus conditions, the data were reworked in order to control for other variables. The additional variables being controlled for were Percent English Learners (EL_PCT), Average Income (AVG_INC), Percent on the Meal Plan (MEAL_PCT), and Average Expenditures per Student (EXP_STU). These variables were chosen for the reason that they are all present in both the California and the Massachusetts samples, which allows for easy comparison. In addition, those variables seemed to cover a few different aspects of the condition of the school and were selected to provide a good amount of information about the school.

In order to take the additional variables into account, the data were sorted into groups of schools that were similar to each other in nearly every way, except for STR. The goal of this concept is to isolate the effect of class size on a sample of very similar schools, thus approaching ceteris paribus conditions. To determine similarity between schools, each variable was classified into bins of certain ranges.

The variable classifications for California are as follows:

- **MEAL_PCT (%) and EL_PCT (%)**: 0 – 25, 25 –50, 50 – 75, 75 – 100
- **EXP_STU**: Less than 4500, 4500 – 5000, 5000 – 5500, 5500 – 6000, over 6000
- **AVG_INC**: Less than 6, 6 – 10, 10 – 14, 14 – 18, and over 18

The variable classifications for Massachusetts are as follows:

- **MEAL_PCT (%) and EL_PCT (%)**: 0 – 25, 25 – 50, 50 – 75, 75 – 100
- **EXP_STU**: Less than 4000, 4000 – 5000, 5000 – 7000, 7000 – 8000, over 8000
- **AVG_INC**: Less than 12, 12 – 16, 16 – 20, 20 – 24, 24 – 32, over 32.

The above ranges were chosen using information like the mean and variance of each variable to capture as much relevant information in the groups as possible. Once each school had been classified, schools that appeared very similar were grouped together in clusters. "Similar" in this context means that schools within a single group only differed from each other by at most two variables, and those variables differed by at most one class number. While each group was not entirely homogenous, on the whole there was not much variation within groups.

Now that schools were organized in such a way that Ceteris Paribus conditions were more fully satisfied than in the original sample, the effect of small classes could be isolated from the other variables. While this method provides a better look at the true effect of class size, no individual group sample had a very large sample size, which introduces potential problems.

F. Randomized Experiment (Causal Effect and Selection Bias): Tennessee Study

In the 1980s, the state of Tennessee conducted a large, randomized controlled experiment to figure out whether reducing class size was an adequate way to improve education in elementary schools. A true randomized experiment is one which is free of potential biases and is representative of the whole population. It has two characteristics- it is free from selection bias, and if so it establishes causal effect.

The study of Tennessee was termed "Project STAR" (Student-Teacher Achievement Ratio), and it remains to be one of the largest random experiments ever conducted in this field of research. After the state invested millions of dollars in this study the researchers concluded that a reduction in class size did indeed result in an improvement in test scores of the students. This experiment was a randomized one since samples were selected as random, and there was no selection bias in the experiment, meaning no particular group benefitted more from an intervention that the entire population. Since the Tennessee experiment is a randomized controlled experiment, it established "ceteris paribus" by controlling for other variables. It thus suggested causal effect between class size and test scores, meaning that no other attributes were statistically significant enough to influence the performance of students. Appendix B shows the confidence intervals of the difference in small and large classes based on attributes such as gender gap, free lunch, and ethnicity (blacks and whites) for this study. Gender and ethnicity attributes are not statistically significant thus suggesting that they do not influence the test scores at all. The free lunch attribute on the other hand does indeed prove to be statistically significant, possibly suggesting that poverty can have some effect on students' performance. However, since the "STAR project" was a randomized controlled experiment and showed causal effect, it can be suggested that the samples drawn happened to fall in the 5% rejection region for the free lunch attribute.

Overcoming selection bias is crucial to making comparisons while "holding everything else constant". Since the Tennessee experiment collected data with respect to categories like minority-only, majority-only, and mixed –race classes, researchers were able to establish causality by showing that students in smaller class sizes scored higher on tests than those in larger class sizes, and that there were no other attributes influencing the test scores.

Results

A. Test scores, student-teacher ratio and other attributes analysis

Table 1 shows how some classroom variables show varying results in small and large classes. Additionally, it provides information on the significance of the variables, meaning whether they are statistically significant to affect the learning outcomes or not.

Table 1. Test scores and characteristics of small and large classes in California and Massachusetts

	California				Massachusetts			
	Small Class	Large Class	Diff.	95% Confidence Interval of difference**	Small Class	Large Class	Diff.	95% Confidence Interval of difference***
Test Scores	657.35 [19.36]	649.98 [17.85]	7.37 (1.82)	7.37 ±3.57	711.22 [14.08]	698.42 [18.58]	12.81 (3.92)	12.81 ±8.07
EL_PCT	12.53 [16.82]	19.99 [19.28]	-7.46 (1.79)	-7.46 ± 3.52	0.89 [2.48]	2.92 [4.91]	-2.02 (1.02)	-2.02 ± 2.09
AVG_INC	16.33 [8.55]	13.98 [4.68]	2.35 (0.65)	2.35 ±1.28	19.12 [5.92]	15.67 [3.61]	3.45 (0.85)	3.45 ± 1.74
MEAL_PCT	41.63 [27.27]	48.72 [26.47]	-7.09 (2.64)	-7.09 ± 5.18	14.00 [13.32]	26.05 [22.82]	-12.05 (4.75)	-12.05 ±9.79
EXP_STU	5540.32 [670.52]	5014.37 [428.94]	525.94 (53.85)	525.94 ± 105.54	5435.89 [976.95]	4834.21 [813.01]	601.68 (180.04)	601.68 ± 370.81

NOTE:
NA stands for “Not Applicable” for the student-teacher ratio attribute
[] denotes the standard deviation, and () denotes the standard error
**using normal distribution table to calculate the confidence interval because n>100
***using student’s t-distribution to calculate the confidence interval because degrees of freedom were less than 100

A 95% confidence interval for difference in test scores of students in small and large sizes suggests that there is a significant relationship between the two variables (CA: 7.37+ 3.57, df=405; MA: 12.81 + 8.07, df= 26.6) as the null (o; no difference in test scores across class sizes) can be rejected. This implies that in both Massachusetts and California, class size does tend to have a significant impact on test scores.

However, stopping the analysis at this point can very likely give misleading results. Behavior of other variables such as EL_PCT, AVG_INC, MEAL_PCT, and EXP_STU under different class sizes (after controlling for student- teacher ratio) is crucial to take into account as they can influence students' learning outcomes. The confidence interval of the difference in these attributes (Table 1), after controlling for class size, suggests that all of them are statistically significant in both California and Massachusetts. For instance, the confidence interval for the difference in percent of English learners in small classes and large classes does not contain the null in both the states (CA: -7.46 + 3.52, MA: -2.02 + 2.09), and hence suggests that there is variability in the % of English learners across different class sizes. The observation that these variables do in fact provide relevant information, makes them necessary to be considered.

B. *Correlation Matrix Analysis: Test Scores and classroom variables*

A correlation matrix between test scores and classroom variables provides information about how closely variables are related (if any) to the test scores.

Table 2. Correlation between test scores and relevant classroom attributes

Characteristic	**Test Score (CA)**	**Test Score (MA)**
Student Teacher Ratio	-0.226	-0.259
Percent of English Language Learners	-0.644	-0.528
Average Income	0.712	0.623
Percent on Free/Reduced Lunch	-0.869	-0.784
Expenditures per Student	0.191	0.109

Table 2 suggests that there is a negative correlation between student-teacher ratio and test scores for both California and Massachusetts. This means that in both the states, increasing class size most likely leads to a reduction in test scores. However, it is important to note that other classroom variables have a stronger correlation with test scores than student-teacher ratio does, thus pointing out the problem in the initial two variables (STR and test scores) model. For instance, average district income shows a positive strong relation with test scores (CA: 0.712; MA: 0.369), meaning if a student comes from a district that has high income on average, then the test scores will also be high and vice versa. This makes sense because socio-economic status of a family does in fact affect the learning outcomes of students. The correlation is however stronger in California than in Massachusetts.

Similarly, percent of students on Free/ reduced lunch is an indicator of poverty and has the potential to affect the test scores of students. This variable has a very strong negative correlation with test scores (CA: -0.64; MA: -0.53), thus indicating that English learning students tend to perform relatively poorer to non-English learning students. Thus, this significance in correlation between test scores and variables other than student- teacher ratio suggests that for a comprehensive analysis, other attributes must be taken into consideration.

C. *Ceteris Paribus Analysis*

After *ceteris paribus* conditions were set within the clusters, the average treatment effects for each state was found. The average effect of small classes on test scores in California was 2.69, and in Massachusetts it was 3.75. A cursory examination of these averages leads one to the conclusion that test scores are absolutely improved by smaller class sizes, but there are a few caveats that make this conclusion a hasty one.

First, the treatment effect listed above is just an average of all the individual differences of the classified school groups. Of the six school groups in California, only one had a treatment effect significant at the 5% level, and another was significant at a 10% level. The rest of the treatments, including both groups from Massachusetts, did not show any significant difference. A second caveat arises since the variables examined all correlated with test scores, as well as with each other. This results in many of the schools that show similarity in the variables EL_PCT, AVG_INC, MEAL_PCT, and EXP_STU also being similar in class size. This made it difficult to find similar groups that differed only in class size, making the sample size quite small. There were plenty of groupings found that were highly similar, but did not have enough of the treatment group or control group present. This happened especially often in Massachusetts. A second consequence of the strong correlations between variables is that the newly-formed school groups no longer covered the entire population. Schools with characteristics at the fringes of the sample were pinched out, in a sense. A school with an especially high or low EXP_STU value, for example, would likely not be very similar to many other schools, so they were unlikely to be placed in a group for further examination. What this all amounts to is that selection bias still likely played a role in this analysis, as a small sample size and values that do not fully represent the population both tend to lead to more bias.

Taking all of this into consideration, the data in Appendix E suggests that the treatment effect found in California and Massachusetts is not significant, as nearly all of the samples from which the average is taken had treatment effects that were not significant. Despite the potential hindrances of selection bias, the lack of significant results in many of the groups shows a level of consistency that is meaningful.

Conclusion

The simple concept that reducing class size could have a beneficial effect on students' learning outcomes is an enticing option for policymakers, especially when bolstered by statistics that show a strong negative correlation between the two. Though it is a simple idea in nature, it is not financially trivial to pursue. Therefore, establishing a causal relationship between class size and test scores is crucial, to prevent a waste of funds.

This study revealed that when just four additional variables are considered, the correlation that seemed so concrete in a vacuum is now a bit more nebulous. In fact, the other variables examined in this study could prove to be equally good or better targets for policy decisions. Increased expenditures per student, for example, could be looked into as a possible option for improving student learning outcomes, though likely equally as expensive.

The takeaway is clear: there is probably not a great way to improve something as complex as student learning outcomes by focusing on mainly one variable, especially if it fails to produce significant results in a *ceteris paribus* analysis.

Appendix

Appendix A: Re-creating Table 1.2 with Real Data

	Loleta Union Elementary	Sebastopol Union Elementary
Potential Test Score (Class>20)	?	?
Potential Test Score Class<20	?	?
Class < 20? (Y/N)	1	1
Actual Test Score	653.85	667.20
Class Size Effect	?	?

Appendix B: STAR Tennessee Data on variables affecting test scores in small and big classes

	Small Class Diff.	**Large Class Diff.**	**Diff. between small and large class**	**Confidence Interval (95%) of difference in Small Class and Large Class**
Gender (girls and boys)	17.08*	13.89**	3.18*** (9.14)	3.18 ± 17.91
Meal Plan	16.66	35.86	-19.20 (8.43)	-19.20 ± 16.54
Ethnicity	26.44	14.08	12.36 (9.98)	12.36 ± 19.58

Note:
* Difference between girls and boys in small class size
**Difference between girls and boys in large class size
***Difference between the gender gap in small class size and the gap in large class size Other variables values were calculated in a similar way to the gender variable

Appendix C: Lower Triangular Correlation Matrix for California Schools

	Test Scores	STR	EL_PCT	AVG_INC	MEAL_PCT	EXP_STU
Test Scores	1	-	-	-	-	-
STR	-0.2263	1	-	-	-	-
EL_PCT	-0.6441	0.1876	1	-	-	-
AVG_INC	0.7124	-0.2321	-0.3074	1	-	-
MEAL_PCT	-0.8687	0.1352	0.6531	-0.6844	1	-
EXP_STU	-0.5402	0.0682	0.3968	-0.3772	0.4851	1

Appendix D: Lower Triangular Correlation Matrix for Massachusetts Schools

	Test Scores	STR	EL_PCT	AVG_INC	MEAL_PCT	EXP_STU
Test Scores	1	-	-	-	-	-
STR	-0.2585	1	-	-	-	-
EL_PCT	-0.5279	0.1623	1	-	-	-
AVG_INC	0.6234	-0.1566	-0.2380	1	-	-
MEAL_PCT	-0.7842	0.1807	0.6623	-0.5627	1	-

EXP_STU	0.0309	0.0756	-0.0596	-0.0105	-0.0425	1

Appendix E: The treatment effect on test scores within clusters

	California	Massachusetts
	1.23	0.36
	-2.50	7.14
	6.15	
	4.47	
	7.38	
	-0.59	
Average	2.69	3.75

Appendix F: The demographic breakdown of California and Massachusetts Clusters

Sample	EXP_STU	AVG_INC	MEAL_PCT	EL_PCT
CA1	less than 5000	Greater than 14	Q1	Q1
CA2	5000-6000	Greater than14	Q1	Q1
CA3	5000-6000	10 to 18	Q2	Q1
CA4	5000-5500	10 to 18	Q2	Q1
CA5	5000-5500	10 to 14	Q3	Q3
CA6	4500-5500	6 to 14	Q4	Q2

MA1	less than 5000	Less than 12	Q1	Q1
MA2	less than 5000	12 to 16	Q1	Q1

Bibliography

1. Archibald, Sarah. "Narrowing in on educational resources that do affect student achievement." *Peabody Journal of Education* 81, no. 4 (2006): 23-42.
2. Angrist, J. and J. Pische. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton: Princeton University Press, 2014.
3. Barnes, John. "The Correlation Matrix: Simple Tool, Powerful Insights & Clear Priorities." *All Analytics*. Last modified July 16, 2012. http://www.allanalytics.com/author.asp?section_id=1413&doc_id=247352.
4. Felber-Smith, Abigail. "Learning Outcomes and Students Learning English as a Second Language." *Networks* 11, no. 2 (2009).
5. Finn, Jeremy D. "Class-size Reduction in Grades K-3." *School reform proposals: The research evidence* (2002): 27-48.
6. Fredriksson, Peter and Björn Öckert. "Resources and Student Achievement: Evidence from a Swedish Policy Reform." *The Scandinavian Journal of Economics* 110, no. 2 (2008): 277-96.
7. Hanushek, Eric A. and Ludger Woessman. "Education and Economic Growth." *International Encyclopedia of Education* (2010), vol. 2: 245-252.
8. Hanushek, Eric A., Susan E. Mayer, and Paul Peterson. "The Evidence on Class Size." *Earning and learning: how schools matter* (1999): 131-168.
9. Jensen, Eric. "How poverty affects behavior and academic performance." *Teaching with poverty in mind: What being poor does to kids' brains and what schools can do about it* (2009): 13-45.
10. Keller, Gerald. *Statistics for Management and Economics*. Stamford: Cengage Learning, 2014.
11. Mitchell, Douglas E. and Ross E. Mitchell. "The Impact of California's Class Size Reduction Initative on Student Achievement: Detailed Finding from Eight School Districts." Riverside, CA: University of California, California Educational Research Cooperative, 1999.
12. Mosteller, Frederick. "The Tennessee study of class size in the early school grades." *The Future of Children* (1995): 113-127.
13. Stock, J. H., and Watson, M. W. *"Introduction to econometrics."* (2007).