credo

CENTER FOR RESEARCH ON EDUCATION OUTCOMES

Urban Charter School Study Technical Appendix

2015

Table of Contents

Introduction 5
Model Selection 5
Internal Validity 6
External Validity 7
Final Decision 9
VCR Exception for New Orleans 10
Developing the CREDO Model 10
Incorporating Feedback 12
Constructive Feedback and Response 12
Data 14
Defining Urbanity 15
Identifying Urban Regions for Inclusion and Eligible Schools Within Each Region 15
Testing the Model 16
Average VCR Growth by Subgroup - 3 Year Model 16
Comparison of Average Charter Effect by Quartile of Starting Score 17
Issues Associated with Repeated Tests of Statistical Significance 18
Number of Students in Each VCR by Subject – 3 and 5 Year Data Sets 19
Regression Output for National Result 20
References 24

Table of Figures

Figure 1: Overview of Research Design **9**

Table of Tables

Table 1: Average VCR Effect Sizes by Subgroup in 3 Year Model 16
Table 2: Comparison of Average Charter Effect by Quartile of Starting Score - Math 17
Table 3: Comparison of Average Charter Effect by Quartile of Starting Score - Reading 18
Table 4: Number of Students in Each VCR by Subject – 3 and 5 Year Data Sets 19
Table 5: National Regression Output, Overall Models 20
Table 6: National Regression Output, Sub-Population Models 22

Urban Charter School Study Technical Appendix **2015**

Introduction

The technical appendix contains six sections. The first section, "Model Selection", discusses how CREDO chose to use the Virtual Control Record (VCR) technique employed in this paper and the relative benefits and drawbacks to this and other commonly used analytic methods. The second section, "Developing the CREDO Model," explains the development of the CREDO regression model and describes how comparisons are made across different states and testing regimes. This section also explores the feedback CREDO has received on the VCR method since the release of our original National Charter School Study in 2009 and how it has been incorporated into our analytic process. The third section, "Data," discusses how test scores across all states were standardized, as well as the reasons for indicator variable omission where necessary. The fourth section, "Defining Urbanity," discusses how CREDO chose which regions and schools to include in this report. The fifth section includes tests of the robustness of CREDO's modeling specification. The sixth and final section provides full regression output from the primary (aggregated urban region) regressions.

Model Selection

Every researcher attempting to accurately estimate the performance of charter schools must address a series of challenges for their models to best approximate the actual impact of enrollment in a charter school relative to alternative educational options. Two major concerns when attempting to measure the

impact of charter enrollment are the internal and external validity of the modeling approach[1]. These issues, and how CREDO selected its analytic technique to best address them, are discussed in this section.

Internal Validity

The internal validity of an analytic method refers to how well it can eliminate the influence of extraneous factors and isolate the "value add" of attendance in a charter school. To do this, researchers must create a counterfactual to represent the growth that each charter student would have expected had they enrolled in a traditional public school (TPS). Experimental methods provide the most valid counterfactual by exploiting random lotteries held at oversubscribed charter schools. Since the mechanism by which students are "selected in" or "selected out" of a charter school is presumably random, these groups of students will on average be similar in both observed and unobserved characteristics. Estimates of charter effects from lottery studies therefore provide a comparative benchmark to judge the ability of other methods to identify the real charter "value add" for the same sample of students.

Since the release of CREDO's first national report in 2009, there have been multiple comparisons between the results found using the VCR method and both experimental and quasi-experimental methods on the same or similar groups of students. An independent analysis of non-experimental research methods conducted by Mathematica Policy Research found that CREDO's VCR method produced results that were not significantly different from an experimental lottery analysis of charter school performance. The same study also noted that the VCR method produced results that were more consistent with the experimental results than other non-experimental methods, including fixed effects[2]. A recent review of the literature also found that results produced by the VCR method gave very similar results to a lottery study undertaken in New York City[3]. The VCR method was also found to perform as well or better than fixed effects models on the same cohort of students.[4] A potential weakness of the VCR method is that charter and TPS students matched on observable characteristics may nonetheless differ in unobserved ways. If these unobservable differences drive the sorting of students between TPS and charter schools, this could

[1] Betts, J. and Hill, P. et al. (2006). "Key Issues in Studying Charter Schools and Achievement: A Review and Suggestions for National Guidelines." National Charter School Research Project White Paper Series, No. 2.
[2] Forston, K. and Verbitsky-Savitz, N. et al. (2012). "Using an Experimental Evaluation of Charter Schools to Test Whether Nonexperimental Comparison Group Methods Can Replicate Experimental Impact Estimates," NCEE 2012-4019, U.S. Department of Education.
[3] Betts, J. and Tang, Y. (2011) "The Effect of Charter Schools on Student Achievement: A Meta-Analysis of the Literature." National Charter School Research Project.
[4] Davis, D. and Raymond, M. (2012). "Choices for Studying Choice: Assessing Charter School Effectiveness Using Two Quasi-Experimental Methods." Economics of Education Review 31(2): 225-236.

introduce bias into the estimate of charter effect. The similarity between results found using experimental and VCR methods noted above suggests that the impact of these unobserved differences is not very impactful in this context.

For the 2013 National Charter School Study, CREDO compared the results found using the VCR method to the results from a fixed effects estimation on the *same group of students*. These were students that both switched from TPS to charter in the period of analysis and for whom CREDO was able to construct a VCR. Results from both models were found to be generally consistent for the same groups of students, with marginal charter impacts from the fixed effects analysis trending lower than those found using the VCR approach. This is likely the result of two major factors. First, fixed effects analyses only include students that switch between charters and TPS, and these students may not be representative of the charter population as a whole. Second, as noted above, CREDO limited the "head to head" comparison of fixed effects and VCR methods to only students that switched from TPS to charter schools, and excluded students that move from charters to TPS. This was done because the VCR method by its construction only captures students who either switch from TPS to charter or "grow up" charter; if a charter student switches back to TPS they are no longer followed (although they would eligible to become a VCR once enrolled back in a TPS).

To see if limiting the "head to head" comparison to only students that switch from TPS to charter affected our estimates in the National Charter School Study, CREDO reran our comparison of fixed effects and VCR methods, this time including students that switch between the charter and TPS sectors in either direction (as would be the case in a traditional fixed effects estimation). The results for this model were indeed closer to the overall findings for that report.

Similar to the National Charter School Study, CREDO found evidence of a slight downward trend among TPS students included in this analysis of urban charter sectors across the United States. Given that in a fixed effects estimation students act as their own control, the existence of an exogenous downward trend in academic performance will bias downward the estimated impact of charter enrollment in our analysis. As students are more likely to switch from TPS to charters than the reverse, an oversampling of early TPS records will bias up the TPS counterfactual from which marginal charter effects are calculated. For this reason, fixed effects estimation techniques are not utilized in this analysis.

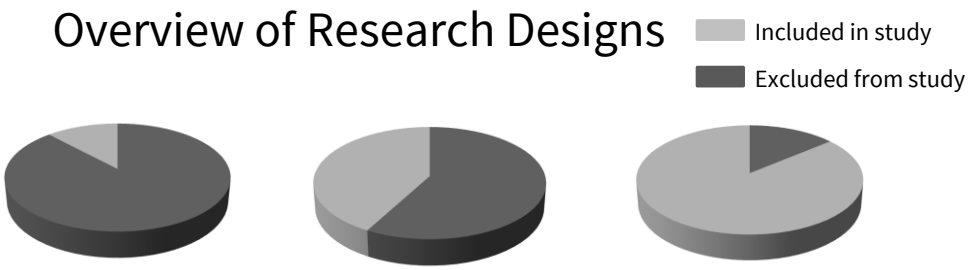
External Validity

A study is considered externally valid if the results can be generalized beyond the specific sample under consideration to a broader population. Lottery studies often have weak external validity due to the fact that they can only derive estimates of charter impact from sufficiently oversubscribed charter schools. To the extent that the average quality of the rest of the charter sector differs from this subset of over-

subscribed schools, results found using lottery analyses generalize weakly to the rest of the charter population. As fixed effects estimation methods work by comparing a student's growth at a charter school to their own prior or subsequent performance at a TPS, the estimate of charter effect can only be calculated for students that switch between charters and TPS in the period of analysis. These "switchers" may not be representative of the rest of the student population (e.g. students who begin their education in charter schools). As a result, the increasing percentage of charter students who "grow up charter" reduces the external validity of fixed effects estimation methods when attempting to generalize to the entire population of charter school students.

The VCR method used by CREDO does not have these limitations to external validity, as all charter students with at least two consecutive test scores are eligible to be included in our study. One characteristic which may lessen the external validity of the VCR method is that the likelihood of a charter student finding a TPS match falls as the student's prior test score (the one on which they are matched) reaches the tail of their states' distributions. However, CREDO's VCR match rate of greater than 80 percent indicates strong external validity remains. An overview of the pros and cons of random assignment, fixed effects and VCR methods is presented in Figure 1 below.

Figure 1: Overview of Research Design



Lottery / Random Assignment	Student Fixed Effects (FE)	Virtual Control Record (VCR)
Unit = Schools	Unit = Students	Unit = Students
Narrow subset of older schools included	Resulting sample biased to middle schools	Majority of students included
Random assignment minimizes selection bias – no guarantee on single draw	Students act as own controls, selection bias minimized	Students matched on fixed characteristics and prior achievement
Uncertain ability to generalize results	Do they accurately reflect all students?	Strong ability to generalize to full population

Final Decision

CREDO concluded that the VCR method provides the best balance between addressing issues of selection bias (internal validity) and using data that is representative of the charter sector as a whole (external validity). Multiple independent confirmations have strengthened CREDO's confidence that the VCR method is at least as internally valid as other quasi-experimental techniques used in the literature, and does not lead to significantly different conclusions than would be the case if we had used experimental methods on the same subset of students. Further, the VCR method maintains high internal validity without compromising external validity. Combined with CREDO's unrivaled data holdings and the VCR method's ability to include the majority of charter students in our estimate of charter effects, we are confident that the results presented in this analysis are the best estimate of the quality of the national urban charter sector to date.

VCR Exception for New Orleans

Since Hurricane Katrina, the New Orleans area has departed from the standard school organization model. For a variety of reasons, the school district in New Orleans decided to move away from the traditional district model and has since converted the vast majority of district schools to charters schools. This situation led to a unique challenge for the VCR process. One aspect of the VCR match is that the pool of potential matches for any charter student is limited to the feeder schools for that student's particular charter school. A feeder school is a traditional public school (TPS) from which the charter school receives students. For example, if charter school J receives students from TPS schools A, B, and D only, then only students from TPS schools A, B, and D will be considered as possible matches for the student from charter J. Even if a student from TPS C was a perfect match, that TPS C student would not be included in the match because he did not come from a charter school J feeder school.

Since there are no longer TPS schools to feed into the New Orleans charter schools, this aspect of the VCR match process had to be modified for New Orleans schools only. For each charter school in New Orleans, we have created a specific state-wide list of Louisiana TPS which have similar student body demographics to that particular charter school. These schools make up our similar schools list for New Orleans charter schools. In the rest of the country, we draw potential student matches from the individual charter school's feeder schools lists. For New Orleans charters, we draw the potential matches from the individual charter school's similar schools list. Other than using a different type of list to locate potential matches, the VCR process for New Orleans is unchanged from other areas of the county. Students are still matched on their individual characteristics to students who are demographically identical to them.

Developing the CREDO Model

After constructing a VCR for each charter student, CREDO then set out to develop a model capable of providing the best estimate of charter impact. The National Charter School Research Project provides a very useful guide to begin the process[5]. First, it is necessary to consider student growth rather than achievement, otherwise controlling for each student's educational history as well as the many observable differences between students that effect their academic achievement is impossible. CREDO's baseline model includes controls for each student's grade, race, gender, socio-economic status (as estimated by eligibility for free or reduced price lunches), special education status, English language learner status and whether the student was held back the previous year. Literature on measuring

[5] Betts, J. and Hill, P. et al. (2006). "Key Issues in Studying Charter Schools and Achievement: A Review and Suggestions for National Guidelines." National Charter School Research Project White Paper Series, No. 2.

educational interventions[6] found that the best estimation techniques must also include controls for baseline test scores. Each student's prior year test score is controlled for in our baseline model. Additional controls are also included for year and period (1st year in the data, 2nd year in the data, etc.). CREDO's baseline model is presented below.

$$\Delta A_{i,t} = \theta A_{i,t-1} + \beta X_{i,t} + \rho Y_t + \sigma S + \gamma C_{i,t} + \varepsilon_{i,t} \quad (1)$$

where the dependent variable is

$$\Delta A_{i,t} = A_{i,t} - A_{i,t-1} \quad (2)$$

And A_{it} is the z-score for student i in period t; $A_{i,t-1}$ is the z-score[7] for student i in period t – 1; X_{it} is a set of control variables for student characteristics and period, Y_t is a year fixed effect, S is a state fixed effect[8]; C is an indicator variable for whether student i attended a charter in period t; and ε is the error term.

In addition to the baseline model above, CREDO explored additional interactions beyond a simple binary to indicate charter enrollment. These included both "double" and "triple" interactions between the charter variable and student characteristics. For example, to differentiate the impact of charter schools by racial group, we estimate models that separate the aggregate charter variable into "charter_black," "charter_hispanic," etc. To further break down the impact of charters by race and poverty, the variables above were split again. For example, black students in charter schools are split further into students that qualify for free and reduced price lunches ("charter_black_poverty") and those that do not ("charter_black_nonpoverty").

[6] Betts, J. and Tang, Y. (2011) "The Effect of Charter Schools on Student Achievement: A Meta-Analysis of the Literature." National Charter School Research Project.
[7] Student z-scores are computed at the grade-by-year level in both reading and math.
[8] Results are ordinally consistent in models run with and without state fixed effects, as well with those found including "state by year" fixed effects.

National aggregate regression results are Errors in Variables regression estimates with standard errors clustered at the school level. For regional regressions, Ordinary Least Squares (OLS) models were used due to significant variation in testing protocols across states and the abnormal test score distribution of students in most urban regions (i.e. urban areas enroll disproportionately lower performing students relative to their state average). In general, regional EIV estimates were consistent with but less conservative than estimates achieved through OLS. The decision was made not to cluster errors at the school level in stratified regional models due to the existence of urban regions that, while containing a substantial number of students in total, nonetheless had a large number of schools with relatively small student bodies (exacerbated by further reduction in effective sample sizes, such as when breaking out charter effects by year). Once the school population eligible for inclusion in stratified regional regressions is reduced further by the elimination of unmatched students and those in untested grades, clustering standard errors at the school level reduces aggregate statistical power to a degree that more than offsets the benefit of estimating standard errors at the school level.

When stratified regional regressions are run with clustered standard errors at the school level, the overall findings are similar, although the ratio of urban regions with positive vs negative marginal effects increases. Specifically, when clustered standard errors are specified, the number of urban regions with significantly lower charter growth in math falls from 11 to 3, while the number of urban regions with significantly greater growth in math falls from 26 to 11. In reading, the number of urban regions with significantly lower charter growth falls from 10 to 3 with clustered standard errors, and the number of urban regions with significantly greater growth falls from 23 to 13.

Incorporating Feedback

CREDO's analytic method has benefited from feedback received by fellow education researchers since the release of our national report in 2009. This feedback covers a broad array of concerns, from potential challenges to the VCR method to problems of estimation and matching protocols. CREDO has found this feedback to be constructive and, even when the particular criticism has turned out to be unfounded in the case of our analysis, it is nonetheless vital to the continuous improvement of our research process and to the scientific method more generally. A discussion of this constructive feedback, and its impact on our research design, fills the rest of this section.

Constructive Feedback and Response

A. After the release of CREDO's first national report in 2009, it was argued that the VCR methodology had the potential to introduce bias into the estimation of charter effect[9].

[9] Hoxby, C. (2009). "A Serious Statistical Mistake in the CREDO Study of Charter Schools." NBER working paper. Available at http://credo.stanford.edu/reports/memo_on_the_credo_study.pdf

Specifically, the concern centered on the fact that student test scores are used both in the calculation of the dependent variable (student growth) and as an independent variable (prior test score). Since charter students are compared to virtual twins, which may include multiple TPS students, there was speculation that the standard error of starting scores for charter students could be significantly larger than for their VCRs, potentially biasing downward the estimated effect of charter enrollment. This is not a valid concern in our analysis, as the standard errors of the starting scores of charter students and their VCRs in period 1 (the year in which they are matched) are not significantly different (as was true in our 2009 and 2013 national reports as well[10]). In fact, standard errors for Charter and VCR starting scores are identical to at least the fourth digit for all major subgroups and for each decile of starting score as well. While this criticism turned out to be invalid, it is nonetheless a theoretically plausible concern and, as a result, CREDO now limits the number of TPS students in each VCR to a maximum of 7 to minimize the possibility of this becoming an issue in the future.

B. Concern was raised that CREDO's decision to allow variation on student's starting scores by up to plus or minus 0.1 standard deviations in the match process may bias the estimate of charter effect[11]. An independent analysis conducted by Mathematica Policy Research found that restricting the variation on starting scores allowed in the match process did not significantly alter the measured impact of charter schools, but it did reduce the proportion of the charter sector that was able to be matched to TPS. Despite this, CREDO believes that this is a potentially valid concern for certain subgroups of charter students whose members lie disproportionately at tails of their state's distributions. For these students, the variance of TPS student's prior year test scores may not be evenly distributed above and below their matched charter students' test scores. To see whether this could bias any of our estimates of charter effect, CREDO tested whether the starting scores of charter students and their VCRs were different in each subgroup. It was found that starting scores are not significantly different for any subgroup analyzed in this report.

C. Analytic approaches that use null hypothesis significance testing (NHST) to determine the presence of relationships between variables can occasionally create the false impression that significant differences exist between two groups of observations when in fact they do not. These false positives, also known as Type 1 errors, are more likely as the number of tests of statistical significance increases. In the construction of CREDO's quality curve, we include not only a charter school's average effect compared to their local environment but also a test of whether

[10] CREDO. (2009) "CREDO Finale to Hoxby's Revised Memorandum." Available at http://credo.stanford.edu/reports/CREDO%20Finale%20to%20Hoxby.pdf

[11] Hoxby, C. (2009). "A Serious Statistical Mistake in the CREDO Study of Charter Schools." NBER working paper. Available at http://credo.stanford.edu/reports/memo_on_the_credo_study.pdf

this effect is significantly different as well. Each of these school breakouts could be considered a separate test of statistical significance[12]. CREDO believes that common corrections for multiple tests of statistical significance can cause more harm than good, and are not well matched to the likely range of charter effect sizes across the country (see "Testing the Model" section of this appendix).

Data

This study built on the methodology used in the 2013 report by creating a pooled set of standardized data from across all states in the study. CREDO combined the data from each state into a single data set in a way that takes the different test measurement scales of each state and turns them into a common set of measures. To do this, CREDO converts each test score into a z-score based on grade-by-year means and standard deviations, which translates each score into a unit of standard deviation. For example, if a student has a z-score equal to zero, this signifies that their test score in that year put them exactly at the 50th percentile in their state, with half of the students taking that test scoring higher and half scoring lower. This transformation allows test scores to be combined across grades and states into a single measure, because each student's growth per year is calculated *relative only to other students in their state and grade in a given year*.

To determine the charter "effect size" for a given subgroup, we compare the growth of each student in that subgroup from the charter sector to the growth of their VCR. For example, if the average black student in a charter school saw their z-score increase from 0 s.d. to 0.1 s.d. (moving from the 50th to the 54th percentile of their state's distribution), while their VCRs saw a z-score increase from 0 s.d. to 0.05 s.d. (moving from the 50th to the 52nd percentile), this would equate to a charter "effect size" of (0.1 – 0.05) = 0.05 s.d.. This is the marginal benefit of attending a charter school for black students on average.

Every state's test also has a level of inaccuracy that cannot be avoided, and this varies not just across states but also for each grade and test score as well. For any given test score, some students will have knowledge and ability greater or less than the score indicates, while for many other students the score will be an accurate reflection of their knowledge at that time. The extent to which a test is capable of accurately reflecting each student's ability is referred to as test reliability. To ensure that the results presented in this paper were robust to differences in reliability of each state's standardized tests, CREDO ran each of our models using STATA's "errors in variables" regressions as well as OLS. Because the charter sector in each state is not necessarily distributed normally across their state's test score distribution,

[12] Mathematica. (2012). "Charter School Performance in New Jersey." What Works Clearinghouse Quick Review. Available at http://ies.ed.gov/ncee/wwc/quickreview.aspx?sid=220

CREDO calculated reliability using standard errors of measurement by grade and score for each state separately.

To avoid over specification among indicator variables for grade and urban region, 5th grade and Columbus were chosen as referent variables for grade and region, respectively (i.e. they were excluded from the regression analysis). 5th grade was chosen for exclusion for multiple reasons. First, we needed to choose a grade that was tested in all states. And second, we didn't want as a reference point any grade with a large number retained students (e.g. 3rd grade). Columbus was chosen for exclusion among region dummies because their marginal region-wide charter effect is closest to the average national charter effect (i.e. the coefficient on Columbus's region fixed effect is closest to 0 in a pooled national regression) for both math and reading. In addition, the average math z score growth rate in Columbus is 0.0005 s.d., while in reading it averages 0.017 s.d.. This eases interpretation of graphical displays of regional means and growth, as states with positive and negative marginal impacts relative to Columbus can be interpreted (roughly) as those above and below average achievement or growth for urban regions across the U.S. for national regressions.

Defining Urbanity

Identifying Urban Regions for Inclusion and Eligible Schools Within Each Region

Conducting an analysis of urban school systems requires a series of decision rules around the selection of urban regions and schools within said regions. CREDO conducted an extensive analysis to identify urban regions for inclusion in this report. Factors considered include total city population, total population of charter students, total number of urban charter students, size of the primary school district(s), and charter market share. After identifying all of the urban regions that rank highly on any of these metrics, and cross referencing this list with available data, a final list of 42 urban regions remained.

The next challenge involves identifying schools for inclusion as city limits, school district boundaries and school addresses do not always align cleanly. For inclusion in this analysis, a school (charter or TPS) must be designated as an urban school by the National Center for Education Statistics and meet at least one of the following criteria: the school is located within the city according to the Common Core of Data, the school is located in the primary school district(s) serving the city of interest, or the school's physical address falls within the city.

Testing the Model

Average VCR Growth by Subgroup - 3 Year Model

By their construction quasi-experimental methods, such as those used in this paper, are comparisons of the growth between charter and TPS students on average. Therefore, a large and positive effect size for a subgroup of charter students could be due to either high levels of growth in the charter sector or due to low levels of growth among the TPS students to which they are being compared (or both). The average effect sizes for each major VCR subgroup below provide a sense of the "yardstick" that the charter sector must reach with each group to achieve a positive marginal effect. Effect sizes by VCR subgroup are found in Table 1 below.

Table 1: Average VCR Effect Sizes by Subgroup in 3 Year Model

Student Group	Reading	Math
Students in Poverty	-0.13	-0.10
ELL Students	-0.32	-0.16
Special Ed Students	-0.33	-0.23
Black Students	-0.24	-0.22
Hispanic Students	-0.12	-0.11
Asian Students	0.07	0.13
Native American Students	-0.16	-0.17
Retained Students	-0.09	-0.004 (not sig)

All results significant at 1% level unless otherwise specified.

Comparison of Average Charter Effect by Quartile of Starting Score

Examining only the average effect of charter enrollment may mask differences in the impact that charter schools have on particular subgroups based on the level of academic preparation of the students within that subgroup. For example, we see below that the positive effect of enrolling in a charter school for a black student in poverty is significantly larger for those who started in the top half of the test score distribution (quartiles 3 & 4) than for those who started in the bottom quarter (quartile 1). Charter effect sizes, stratified by quartile of starting score in period 1, are presented in Tables 2 and 3 below. All effect sizes are significant at the 1% level or greater unless otherwise indicated.

Table 2: Comparison of Average Charter Effect by Quartile of Starting Score - Math

Starting Score in Period 1 Variable	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Charter	.052	.062	.066	.047
Charter Students in Poverty	.043	.038	.039	.035
Charter Ell Students	.058	.048	.053	.041
Charter Special Ed Students	.015 (not sig)	.041	.051	.042
Charter Black Students	.02 (not sig)	.048	.063	.063
Charter Hispanic Students	-.011 (not sig)	.024 (not sig)	.036	.034
Charter Asian Students	.014 (not sig)	.01 (not sig)	.005 (not sig)	.011 (not sig)
Charter Native American Students	-.160	-.092 (not sig)	-.069 (not sig)	-.091
Charter Retained Students	.05 (not sig)	-.005 (not sig)	.016 (not sig)	-.019 (not sig)

All results significant at 1% level unless otherwise specified.

Table 3: Comparison of Average Charter Effect by Quartile of Starting Score - Reading

Starting Score in Period 1 Variable	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Charter	.053	.049	.038	.028
Charter Students in Poverty	.018 (not sig)	.029	.030	.026
Charter Ell Students	.061	.083	.096	.104
Charter Special Ed Students	.012 (not sig)	.012 (not sig)	.033 (not sig)	.069
Charter Black Students	.036	.039	.036	.035
Charter Hispanic Students	.013 (not sig)	.009 (not sig)	.006 (not sig)	.005 (not sig)
Charter Asian Students	-.053 (not sig)	-.011 (not sig)	.002 (not sig)	.006 (not sig)
Charter Native American Students	-.099 (not sig)	-.044 (not sig)	-.082	.017 (not sig)
Charter Retained Students	.045 (not sig)	-.025 (not sig)	-.049 (not sig)	.024 (not sig)

Issues Associated with Repeated Tests of Statistical Significance

CREDO made the decision not to adjust for potential type 1 errors for our school level analyses (such as with a Bonferroni correction) for multiple reasons. First, the Bonferroni correction, and similar procedures that essentially involve lowering the p value threshold for each test of statistical significance, would indeed "correct" the test but for the wrong null hypothesis (i.e. that NONE of the charters are significantly different from their local TPS competitors). For example, if at least one charter school had a p value that met the arbitrarily more stringent threshold, we would then accept the alternative hypothesis that "at least one" of the charter schools had significantly different effects than their TPS competitors. This is not the null hypothesis our school level analysis is designed to test. Instead, we are testing the null hypothesis that each charter school's growth is not significantly different than that of their feeder TPS.

There is a second reason we do not "adjust" our significance tests. In education research the null hypothesis that all of our marginal charter school effect sizes are exactly equal to zero, while necessary

for NHST, is likely not plausible for the purposes of estimating the probability of type 1 errors[13]. In addition, relatively small effect sizes (such as those found in many educational interventions) are further reason to be cautious about reducing the power of one's analysis and deliberately increasing the risk of a type 2 error as a result (not finding a significant difference where one exists).

Number of Students in Each VCR by Subject – 3 and 5 Year Data Sets

In Table 4 below, we report the average number of TPS students that make up each charter student's VCR. This is provided for both the 3 and 5 year data sets. As was shown above, the fact that each VCR record contains multiple TPS students does not affect our ability to accurately estimate the effect of charter school enrollment. The decision to use multiple TPS records in a charter's VCR was based on the desire to get the fairest comparison between a charter student's growth and the growth they could have expected in their alternative TPS environment. CREDO believes that allowing up to 7 TPS matches per charter student provides the best balance between constructing a fair TPS comparison set for our charter students and maintaining the ability to accurately estimate the real "value add" of enrollment in charter schools.

Table 4: Number of Students in Each VCR by Subject – 3 and 5 Year Data Sets

Number of Students in Each VCR by Subject	Mean	Median	SD
Reading – 3 year	5.14	7	2.18
Reading – 5 year	5.17	7	2.17
Math – 3 year	4.96	6	2.23
Math – 5 year	4.99	6	2.22

[13] Gelman, A. et al. (2012). "Why We (Usually) Don't Have To Worry About Multiple Comparisons," Journal of Research on Educational Effectiveness, 5: 189-211.

Regression Output for National Result

Table 5: National Regression Output, Overall Models

Variable	Reading		Math	
	Coefficient	SE	Coefficient	SE
Charter Student	0.039**	(0.001)	0.055**	(0.001)
Starting Score	-0.359**	(0.001)	-0.315**	(0.001)
Black	-0.213**	(0.001)	-0.170**	(0.001)
Hispanic	-0.108**	(0.001)	-0.070**	(0.001)
Asian or Pacific Islander	0.080**	(0.002)	0.156**	(0.002)
Native American	-0.170**	(0.009)	-0.193**	(0.010)
Multi-Ethnic	-0.050**	(0.005)	-0.046**	(0.006)
Is Special Ed	-0.320**	(0.002)	-0.224**	(0.002)
Is English Learner	-0.279**	(0.002)	-0.139**	(0.002)
Is In Poverty	-0.112**	(0.001)	-0.084**	(0.001)
Repeated Grade	-0.088**	(0.003)	0.002	(-0.003)
grade_01	0.621**	(0.139)	0.570**	(0.118)
grade_02	0.128**	(0.010)	0.090**	(0.011)
grade_03	-0.008**	(0.002)	0.004	(-0.002)
grade_04	0.021**	(0.001)	0.003*	(0.001)
grade_06	0.014**	(0.001)	-0.001	(-0.001)
grade_07	0.041**	(0.001)	0.037**	(0.001)
grade_08	0.022**	(0.001)	0.054**	(0.001)
grade_09	0.054**	(0.002)	-0.074**	(0.002)
grade_10	-0.005**	(0.002)	-0.204**	(0.002)
grade_11	-0.042**	(0.002)	-0.266**	(0.002)
grade_12	-1.991**	(0.005)	-0.572**	(0.005)
year_2009	0.006**	(0.001)	0.018**	(0.001)
year_2010	-0.021**	(0.001)	-0.025**	(0.001)

Variable	Reading		Math	
	Coefficient	SE	Coefficient	SE
period_2	0.040**	(0.001)	0.027**	(0.001)
period_3	0.056**	(0.001)	0.050**	(0.001)
period_4	0.062**	(0.002)	0.035**	(0.002)
period_5	0.086**	(0.003)	0.057**	(0.003)
Constant	0.142**	(0.002)	0.079**	(0.002)
Observations	2,037,019		1,965,819	
Adjusted R-squared	0.229		0.181	

**significant at 5%; ** significant at 1 % level*

Table 6: National Regression Output, Sub-Population Models

Variable Label	Reading		Math	
	Coefficient	SE	Coefficient	SE
Starting score	-0.359**	(0.001)	-0.316**	(0.001)
Charter Black	-0.203**	(0.002)	-0.166**	(0.002)
TPS Black	-0.239**	(0.002)	-0.217**	(0.002)
Charter Hispanic	-0.112**	(0.002)	-0.077**	(0.002)
TPS Hispanic	-0.120**	(0.002)	-0.105**	(0.002)
Charter Asian or Pacific Islander	0.072**	(0.003)	0.141**	(0.004)
TPS Asian or Pacific Islander	0.071**	(0.003)	0.129**	(0.004)
Charter Native American	-0.194**	(0.013)	-0.263**	(0.014)
TPS Native American	-0.161**	(0.013)	-0.166**	(0.014)
Charter White	-0.021**	(0.002)	-0.047**	(0.002)
Charter – Special Ed	-0.312**	(0.002)	-0.217**	(0.002)
TPS – Special Ed	-0.329**	(0.002)	-0.230**	(0.002)
Charter – English Learner	-0.244**	(0.002)	-0.118**	(0.002)
TPS – English Learner	-0.315**	(0.002)	-0.159**	(0.002)
Charter – in Poverty	-0.100**	(0.001)	-0.068**	(0.001)
TPS – in Poverty	-0.125**	(0.001)	-0.101**	(0.001)
Charter – Repeated Grade	-0.084**	(0.004)	0.008*	(0.004)
TPS – Repeated Grade	-0.092**	(0.004)	-0.004	(-0.004)
grade_01	0.620**	(0.139)	0.570**	(0.118)
grade_02	0.128**	(0.010)	0.089**	(0.011)
grade_03	-0.008**	(0.002)	0.003	(-0.002)
grade_04	0.021**	(0.001)	0.003*	(0.001)
grade_06	0.014**	(0.001)	-0.001	(0.001)
grade_07	0.041**	(0.001)	0.037**	(0.001)
grade_08	0.021**	(0.001)	0.054**	(0.001)

Variable Label	Reading		Math	
	Coefficient	SE	Coefficient	SE
grade_09	0.054**	(0.002)	-0.074**	(0.002)
grade_10	-0.005**	(0.002)	-0.204**	(0.002)
grade_11	-0.042**	(0.002)	-0.266**	(0.002)
grade_12	-1.991**	(0.005)	-0.573**	(0.005)
year_2009	0.006**	(0.001)	0.018**	(0.001)
year_2010	-0.021**	(0.001)	-0.025**	(0.001)
period_2	0.040**	(0.001)	0.027**	(0.001)
period_3	0.056**	(0.001)	0.050**	(0.001)
period_4	0.062**	(0.002)	0.035**	(0.002)
period_5	0.086**	(0.003)	0.057**	(0.003)
Constant	0.171**	(0.002)	0.128**	(0.002)
Observations	2,037,019		1,965,819	
Adjusted R-squared	0.229		0.181	

**significant at 5%; ** significant at 1 % level*

References

Betts, J. and Hill, P. et al. (2006). “Key Issues in Studying Charter Schools and Achievement: A Review and Suggestions for National Guidelines.” National Charter School Research Project White Paper Series, No. 2.

Betts, J. and Tang, Y. (2011) “The Effect of Charter Schools on Student Achievement: A Meta-Analysis of the Literature.“ National Charter School Research Project.

Borjas, G. (1980). “The Relationship Between Wages and Weekly Hours of Work: The Role of Division Bias.” Journal of Human Resources, Vol. 15(3), (pp. 409-423).

CREDO. (2009) “CREDO Finale to Hoxby’s Revised Memorandum.” Available at: http://credo.stanford.edu/reports/CREDO%20Finale%20to%20Hoxby.pdf

Davis, D. and Raymond, M. (2012). “Choices for Studying Choice: Assessing Charter School Effectiveness Using Two Quasi-Experimental Methods.” Economics of Education Review 31(2): 225-236.

Forston, K. and Verbitsky-Savitz, N. et al. (2012). “Using an Experimental Evaluation of Charter Schools to Test Whether Nonexperimental Comparison Group Methods Can Replicate Experimental Impact Estimates,” NCEE 2012-4019, U.S. Department of Education.

Hoxby, C. (2009). “A Serious Statistical Mistake in the CREDO Study of Charter Schools.” NBER working paper. Available at: http://credo.stanford.edu/reports/memo_on_the_credo_study.pdf

Kutner, M. et al. (2004). “Applied Linear Regression Models, 4th edition, McGraw-Hill Irwin.

Mathematica. (2012). “Charter School Performance in New Jersey.” What Works Clearinghouse Quick Review. Available at: http://ies.ed.gov/ncee/wwc/quickreview.aspx?sid=220