credo

CENTER FOR RESEARCH ON EDUCATION OUTCOMES

National Charter School Study
Technical Appendix
2013

Table of Contents

Table of Figures

Table of Tables

National Charter School Study Technical Appendix **2013**

Introduction

The technical appendix contains 8 sections. The first section, "Model Selection", discusses how CREDO chose to use the Virtual Control Record 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 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 previous national report in 2009 and how it has been incorporated into our analytic process. The third section, "Data," discusses how test scores across 27 states were standardized, as well as the reasons for dummy variable omission where necessary. The fourth section compares charter students that were able to find matches to those that were not. The fifth section includes tests of comparability between charter records and their VCRs. The sixth section includes tests of the robustness of CREDO's modeling specification. The seventh section contains full regression output from the primary 16 (original cohort) and 27 (all in) state regressions. The final section contains a point by point refutation of critiques presented by the Center for Education Reform, a charter advocacy organization.

Model Selection

Every researcher attempting to accurately measure the performance of charter schools must addresses a series of challenges in order for their models to reflect reality. 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 TPS. Experimental methods can 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 can therefore provide a comparative benchmark to judge the ability of other methods to identify the real charter "value add" in the same sample of students.

Since the release of CREDO's 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 introduce bias into the estimate of charter effect The similarity between

[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.

results found using experimental and VCR methods noted above suggests that the impact of these unobserved differences is not very impactful in this context.

In addition to the confirmations of the internal validity of the VCR approach referenced above, CREDO also compared the results found using the VCR method to the results of 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; overall charter effects are presented using each method in table 1 below.

There are two likely reasons that results found for these students are lower than those in the main body of this report. First, this analysis could only be done on students that switch between charters and TPS, and these students may not be representative of the charter population as a whole. In support of this possibility, CREDO found suggestive evidence that charter-bound TPS students were on an accelerating negative trend in the two years before switching to a charter school, and these students had different starting achievement than other students in our analysis. 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 in our analysis switches back to TPS they are no longer followed.

To see if limiting our "head to head" comparison to only students that switch from TPS charter could be affecting our estimation, 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 from the paper. This is likely due in part to a slight downward trend among TPS achievement in our data, (also noted among VCRs in the body of the report). This implies that the exclusion of TPS records from later years, or rather the over sampling of early observations in TPS, biases down the estimated charter effect by biasing up the TPS counterfactual.

Table 1: Comparison of Charter Effects Using Fixed Effects and VCR Methods on the Same Students

Method Used	Only TPS to Charter "Switchers"		Either TPS to Charter *or* Charter to TPS "Switchers"	
	Reading	Math	Reading	Math
Fixed Effects	-0.035	-0.059	-0.028	-0.046
VCR	-0.025	-0.039	-0.017	-0.029

All results are significant at the 1% level.

External Validity

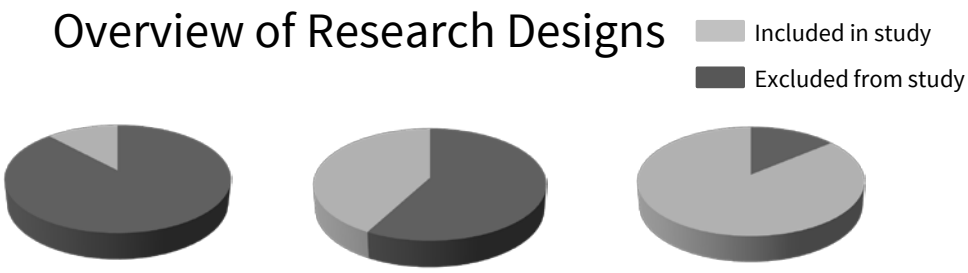
A study is considered externally valid if the results can be generalized beyond the specific sample under consideration to a broader population. Lottery studies often have weak external validity due to the fact that they can only include sufficiently oversubscribed charter schools. To the extent that 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. Another commonly used analytic technique is the fixed effects model. 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 therefore only be calculated for students that switch between charters and TPS in the period of analysis. As these "switchers" are seen not to be representative of the rest of the student population (e.g. students who begin their education in charter schools), this reduces the external validity of fixed effects estimation methods.

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. Further, our data holdings include student-level records from states that enroll over 95 percent of the charter students in the country. Of these students, 44 percent are enrolled in tested grades and have growth scores.

One characteristic that 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. This notwithstanding, CREDO's VCR match rate for tested students is 85 percent, which indicates strong external validity remains. These numbers give the results of this national report a level of validity as strong as any charter study to date.

An overview of the exclusion criteria for experimental, 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 decided 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. 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 measure of the quality of the national charter sector to date.

Developing the CREDO Model

After constructing a VCR for each charter student, CREDO then set out to develop a model capable of providing a fair measure 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, free or reduced price lunch status, special education status, English language learner status and whether they were held back the previous year. Literature on measuring 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 state, year[7] and period (1st year in charter, 2nd year in charter, 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_{it-1} is the z-score 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; C is an indicator variable for whether student i attended a charter in period t; and ε is the error term. Errors are clustered around charters schools and their feeder patterns as well.

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

[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.

[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] State-by-Year fixed effects were modeled to see if changes in state level policy over our period of analysis may impact the results, but results were identical to the specification above.

the charter variable and student characteristics. For example, to identify the impact of charter schools on different racial groups, we estimate models that break the 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").

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 falls into two basic categories and these categories have had very different levels of influence on our research design. The first type of feedback covers a broad array of concerns, from potential challenges with 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.

A number of criticisms of the 2009 national study concerning data quality, data sources, study design and CREDO's general ability to conduct research were raised by the charter advocacy group Center for Education Reform. The list of critiques presented on their website contains a series of basic misunderstandings about CREDO's methods. For the interested reader, CREDO provides a point by point refutation of these critiques in Table 10.

Constructive Feedback and Response

A. After the release of CREDO's previous national report in 2009, it was argued that the VCR methodology had the potential to introduce bias into the estimation of charter effect[8]. 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

[8] 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

estimated effect of charter enrollment[9]. As can be seen in Figures 6 and 7 below, 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 national report 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 (see Figures 4 & 5 below).

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 this effect is significantly different as well. Each of these school breakouts could be

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

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 the technical appendix).

Data

This study built on the methodology used in the 2009 report by creating a pooled set of standardized data from across the 27 states in the study. CREDO combined the data from 27 states into a single data set in a way that takes the different test measurement scales in each state and turns them into a common set of measures. To do this, CREDO converts each test score into a z-score, 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 states into a single measure, because each student's growth per year is calculated *relative only to other students in their state and grade*.

To determine the charter "effect size" for a given subgroup, we compare the growth of that subgroup in 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 its' 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. Because the charter sector in each state is not necessarily distributed normally across their state's test score distribution, CREDO calculated reliability using standard errors of measurement by grade and score for each state separately. While our previous national report in 2009 did not control for the reliability of test scores and used robust rather than clustered errors, results are compared between "errors in variables" and

[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

ordinary least squares regressions (with both robust and clustered standard errors) for the current analysis and are found to be comparable (see Table 5 below), making comparisons of effect sizes between the methods valid.

To avoid over specification among the indicator variables for grade and state, 5th grade and New Mexico were chosen as referent variables for grade and state, respectively (i.e. they are 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). New Mexico was chosen for exclusion among state dummies because their marginal state-wide charter effect is closest to the average national charter effect (i.e. the coefficient on NM's state fixed effect is closest to 0 in a pooled national regression) for both math and reading.

Unmatched v. Matched Students

Comparison of Starting Scores of Matched and Unmatched Students

Although the VCR method used in this report provides matches for 85% of the charter students in our data set, it is important to identify ways in which unmatched students may differ with those included in the analysis. The ability to extrapolate findings from a particular sample to the broader population is referred to as external validity (discussed above). In the case of this analysis, CREDO's sample encompasses a large proportion of the entire population of charter students across the country, but as can be seen below, unmatched charter students do differ from their matched counterparts. We see that the test scores of matched charter students are significantly higher than for unmatched students in both math and reading in the year in which they were matched (period 1). This is because charter students at the very low and high end of the test score distribution have more trouble finding matches in TPS. The fact that our data represent 95% of all charter students in the country makes us confident that estimates are highly aligned with actual population values, although we are uncertain to what extent our results apply to students without matches.

Figure 2: Comparison of Starting Reading Scores of Matched and Unmatched Charter Students

```
Ttest  z_origin     charter matched and unmatched    period1
Yes = matched, No = unmatched

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      No |  267011    -.479623     .002331    1.204514   -.4841917   -.4750542
     Yes | 1296645   -.0512084    .0008305     .945699   -.0528362   -.0495807
---------+--------------------------------------------------------------------
combined | 1563656   -.1243648    .0008058    1.007653   -.1259442   -.1227854
---------+--------------------------------------------------------------------
    diff |           -.4284146    .0021139               -.4325577   -.4242715
------------------------------------------------------------------------------
    diff = mean(No) - mean(Yes)                                   t = -2.0e
Ho: diff = 0                                     degrees of freedom =  1.6e

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Figure 3: Comparison of Starting Math Scores of Matched and Unmatched Charter Students

```
Ttest  z_origin     charter matched and unmatched    period1
Yes = matched, No = unmatched

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      No |  288179   -.4249106    .0022219    1.192786   -.4292655   -.4205557
     Yes | 1240501   -.1051811      .00084    .9355975   -.1068275   -.1035347
---------+--------------------------------------------------------------------
combined | 1528680   -.1654549    .0008064    .9970821   -.1670355   -.1638743
---------+--------------------------------------------------------------------
    diff |           -.3197295    .0020456               -.3237388   -.3157203
------------------------------------------------------------------------------
    diff = mean(No) - mean(Yes)                                   t = -1.6e
Ho: diff = 0                                     degrees of freedom =  1.5e

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Comparability of Charter Records & Their VCRs

Comparison of Starting Scores between Charter Students and Their VCRs

For the comparison of charter students to their VCRs to be a truly "apples to apples" comparison, their starting scores in the year in which they were matched (period 1) should be equal. Otherwise, we cannot be sure that charters and their VCRs enter the year of analysis with equivalent educational endowments. Below we find that there is not a significant difference between the starting scores of charter students and their VCRs. Starting scores for every demographic subgroup (e.g. black, ell, poverty), and interaction (e.g. black charter students in poverty, black charter students not in poverty) were found to be insignificantly different from one another, ensuring that charter students are being matched to TPS students with the same initial level of academic achievement.

Figure 4: Comparison of Starting Reading Scores Between Charter Students and Their VCRs

```
Ttest        z_origin     charter and VCR     Period1

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     TPS | 1296645    -.051082    .0008295    .9445252   -.0527077   -.0494563
 Charter | 1296645   -.0512084    .0008305     .945699   -.0528362   -.0495807
---------+--------------------------------------------------------------------
combined | 2593290   -.0511452    .0005869    .9451121   -.0522955   -.0499949
---------+--------------------------------------------------------------------
    diff |            .0001264    .0011738               -.0021742     .002427
------------------------------------------------------------------------------
    diff = mean(TPS) - mean(Charter)                              t =   0.1077
Ho: diff = 0                                     degrees of freedom =  2.6e

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.5429         Pr(|T| > |t|) = 0.9142          Pr(T > t) = 0.4571
```

Figure 5: Comparison of Starting Math Scores Between Charter Students and their VCRs

```
Ttest        z_origin     charter and VCR     Period1

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     TPS | 1240501   -.1049147    .0008387      .93407   -.1065584    -.103271
 Charter | 1240501   -.1051811      .00084    .9355975   -.1068275   -.1035347
---------+--------------------------------------------------------------------
combined | 2481002   -.1050479    .0005935    .9348338   -.1062111   -.1038847
---------+--------------------------------------------------------------------
    diff |            .0002664     .001187               -.0020601    .0025929
------------------------------------------------------------------------------
    diff = mean(TPS) - mean(Charter)                              t =   0.2244
Ho: diff = 0                                     degrees of freedom =  2.5e

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
Pr(T < t) = 0.5888         Pr(|T| > |t|) = 0.8224          Pr(T > t) = 0.4112
```

Average VCR Growth by Subgroup in 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. For example, for the charter sector to have a positive marginal effect (charter growth – VCR growth) for special education students in math, they need only achieve yearly growth that is greater than 0.20 standard deviations below their state's average growth. However, for the charter sector to have a positive marginal effect for Asian students in math, charter schools must achieve an average rate of growth 0.14 standard deviations higher than their state's average growth. Effect sizes by VCR subgroup are found in Table 2 below.

Table 2: Average VCR Effect Sizes by Subgroup in 3 Year Model

Student Group	Reading	Math
Students in Poverty	-0.10	-0.09
ELL Students	-0.21	-0.10
Special Ed Students	-0.25	-0.20
Black Students	-0.14	-0.16
Hispanic Students	-0.06	-0.06
Asian Students	0.08	0.14
Native American Students	-0.11	-0.11
Retained Students	-0.09	0.001 (not sig)

All results significant at 1% level unless otherwise specified.

Comparison of Starting Score Distribution in Period 1 (3 Year Models)

Figure 6: Test of Starting Scores in Reading for Charters and VCRs in Period 1

```
Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     TPS |  737440   -.0700773    .0011075    .9510166   -.0722478   -.0679067
 Charter |  737440   -.0702477    .0011088    .9521701   -.0724209   -.0680745
---------+--------------------------------------------------------------------
combined | 1474880   -.0701625    .0007836    .9515932   -.0716982   -.0686267
---------+--------------------------------------------------------------------
    diff |            .0001704    .0015671               -.0029011    .0032419
------------------------------------------------------------------------------
    diff = mean(TPS) - mean(Charter)                              t =   0.1087
Ho: diff = 0                                     degrees of freedom =  1.5e+06

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.5433         Pr(|T| > |t|) = 0.9134          Pr(T > t) = 0.4567
```

Figure 7: Test of Starting Scores in Math for Charters and VCRs in Period 1

```
Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     TPS |  705299   -.1152758    .0011169    .9379826   -.1174649   -.1130867
 Charter |  705299   -.1155488    .0011186    .9394397   -.1177413   -.1133564
---------+--------------------------------------------------------------------
combined | 1410598   -.1154123    .0007904    .9387111   -.1169614   -.1138632
---------+--------------------------------------------------------------------
    diff |             .000273    .0015807               -.0028252    .0033712
------------------------------------------------------------------------------
    diff = mean(TPS) - mean(Charter)                              t =   0.1727
Ho: diff = 0                                     degrees of freedom =  1.4e+06
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.5686         Pr(|T| > |t|) = 0.8629          Pr(T > t) = 0.4314
```

Testing the Model

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 based on five periods, stratified by quartile of starting score in period 1, are presented in Tables 3 and 4 below. All effect sizes are significant at the 1% level or greater unless otherwise indicated.

Table 3: 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	-.01 (not sig)	.01	.007	.007
Charter Students in Poverty	.034	.021	.021	.014
Charter Ell Students	.066	.056	.074	.092
Charter Special Ed Students	.031	-.004 (not sig)	.018	.046
Charter Black Students	-.01	.018	.019	.024
Charter Black Students in Poverty	.027	.041	.045	.045
Charter Hispanic Students	-.052	-.007 (not sig)	-.007 (not sig)	.001 (not sig)
Charter Hispanic Students in Poverty	-.01	.018	.017	.016
Charter Asian Students	-.11	-.033	-.016 (not sig)	-.006 (not sig)
Charter Native American Students	-.090	-.043	-.051	-.015 (not sig)
Charter Retained Students	.032	.017 (not sig)	-.021 (not sig)	.045 (not sig)

All results significant at 1% level unless otherwise specified.

Table 4: 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	.014	.011	.006 (not sig)	.007
Charter Students in Poverty	.027	.023	.018	.015
Charter Ell Students	.054	.056	.082	.089
Charter Special Ed Students	.015	-.004 (not sig)	.022	.046
Charter Black Students	.002 (not sig)	.017	.021	.024
Charter Black Students in Poverty	.031	.044	.042	.047
Charter Hispanic Students	-.023	-.008 (not sig)	-.004 (not sig)	.001 (not sig)
Charter Hispanic Students in Poverty	.007	.019	.017	.017
Charter Asian Students	-.079	-.030	-.015 (not sig)	-.006 (not sig)
Charter Native American Students	-.084	-.041	-.057	-.010 (not sig)
Charter Retained Students	.067	.008 (not sig)	-.014 (not sig)	.048 (not sig)

Comparison of Regression Results with and without Controlling for Reliability

Standardized tests, such as those used to create the rates of student growth in this report, are imperfect measures of student achievement. Furthermore, the level of precision varies based on a student's state, grade and where they fall in their state's distribution. CREDO's analysis does not explicitly control for the reliability of each state's test by grade and score. To ensure that our results are robust to test score reliability, and that comparisons made between the results from our 2009 report and this analysis are valid, we conducted a parallel analysis for our 5-year models with and without controlling for test score reliability. Given that the 2009 report was conducted using robust rather than clustered standard errors, these robust OLS models were run as well on the current data. As can be seen below, differences between effect sizes found in the models that control for reliability (EIVREG) and those that don't (ROBUST OLS & CLUSTERED OLS) are trivial enough that 1) comparisons between the 2009 and 2013 results are valid and 2) coefficients are not affected substantially by variations in test score reliability. Effect size differences between the two analyses for each subgroup are generally similar, although the clustered OLS model does lose significance in at least one subject (due to significantly shrunken sample sizes). The top line effect sizes for major subgroups are presented in Table 5 below.

Table 5: Comparison of Regression Results with and without Controlling for Reliability/Clustering

Student Group	Reading	Math
EIVREG	.008	-.008
Robust OLS	.007	-.008
Clustered OLS	.007	-.008 (not sig)

Note: Effect sizes are significant at the 1% level unless otherwise noted.

Variance Inflation Factor (VIF) Test for Multicollinearity

Multicollinearity occurs when there is a linear relationship between two explanatory variables. If two explanatory variables are highly correlated, this can distort the estimated impact of each variable on the dependent variable (in our case, student growth). Multicollinearity is a generally considered a concern if the result of a VIF test is 10 or greater[13]. State level fixed effects, used to control for differences in education environment by state, were found to have VIF greater than 10 for the largest states in our analysis. To ensure our results were not significantly affected by multicollinearity, regressions were run without fixed effects for these states (adding these states to the "baseline" by which other state effects are measured). As you can see below, coefficients on each variable of interest do not appear to be strongly influenced by the inclusion of large state fixed effects. Note that period, year, grade and state fixed effects are included in models below, but the modeling output presented here is restricted to the key independent variables of interest. Four types of output are presented below for both math and reading: 1. Regression *including* all state fixed effects, 2. VIF test for this model, 3. The same regression above but *without* state fixed effects with VIF>10, 4. VIF test for this model.

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

Figure 8: National Regression Including All State Fixed Effects for Math

```
Linear regression                                      Number of obs = 4850274
                                                       F( 56,4850217) =11414.72
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1595
                                                       Root MSE      =  .51706

------------------------------------------------------------------------------
             |               Robust
   grz_state |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    z_origin |  -.2654484   .0003686  -720.19   0.000    -.2661708    -.264726
     charter |  -.0082989   .0004696   -17.67   0.000    -.0092193   -.0073786
       lunch |  -.0779178   .0005656  -137.75   0.000    -.0790265   -.0768092
         ell |  -.0851693   .0012306   -69.21   0.000    -.0875812   -.0827574
        sped |  -.1944165   .0011721  -165.87   0.000    -.1967138   -.1921192
    retained |  -.0139277    .002506    -5.56   0.000    -.0188394    -.009016
    re_black |   -.120493   .0007345  -164.05   0.000    -.1219326   -.1190534
     re_hisp |  -.0389655   .0006979   -55.83   0.000    -.0403333   -.0375976
  re_asianpi |   .1470946    .001415   103.96   0.000     .1443212    .1498679
  re_nativam |  -.1200225    .004294   -27.95   0.000    -.1284386   -.1116063
    re_multi |  -.0224532   .0024108    -9.31   0.000    -.0271784    -.017728
       _cons |   .0841887   .0030416    27.68   0.000     .0782273    .0901501
```

Figure 9: VIF Test for Math Regression with All State Fixed Effects

```
    Variable |       VIF       1/VIF
-------------+----------------------
          CA |     37.52    0.026651
          TX |     21.15    0.047289
          FL |     17.77    0.056273
          MI |     13.21    0.075687
***Excluded for space***
    re_multi |      1.03    0.973296
  re_nativam |      1.02    0.980855
    grade_01 |      1.00    0.995996
     charter |      1.00    0.999992
-------------+----------------------
    Mean VIF |      4.09
```

Figure 10: National Regression Excluding State Fixed Effects with VIF > 10 for Math

```
Linear regression                                      Number of obs = 4850274
                                                       F( 52,4850221) =12071.32
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1578
                                                       Root MSE      =  .51758

------------------------------------------------------------------------------
             |               Robust
   grz_state |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    z_origin |  -.2638365   .0003649  -723.02   0.000    -.2645517   -.2631213
     charter |  -.0083062     .00047   -17.67   0.000    -.0092275   -.0073849
       lunch |  -.0775574   .0005634  -137.66   0.000    -.0786616   -.0764532
         ell |  -.0963679   .0012205   -78.96   0.000    -.0987601   -.0939758
        sped |  -.1888716   .0011716  -161.20   0.000    -.1911679   -.1865752
    retained |  -.0011532   .0025062    -0.46   0.645    -.0060652    .0037589
    re_black |  -.1192097   .0007252  -164.38   0.000    -.1206311   -.1177883
     re_hisp |  -.0321946   .0006773   -47.53   0.000    -.0335222   -.0308671
  re_asianpi |   .1413233   .0014067   100.46   0.000     .1385662    .1440805
  re_nativam |  -.1235291   .0042812   -28.85   0.000    -.1319201    -.115138
    re_multi |   -.024252   .0024125   -10.05   0.000    -.0289803   -.0195236
       _cons |    .074914   .0009433    79.42   0.000     .0730652    .0767629
------------------------------------------------------------------------------
```

Figure 11: VIF Test for Math Regression without Four Large State Dummies

```
    Variable |       VIF       1/VIF
-------------+----------------------
   year_2009 |      2.61    0.382532
   year_2008 |      2.44    0.409053
   year_2007 |      2.30    0.434717
   year_2006 |      2.15    0.465606
****Excluded for Space****
  re_nativam |      1.01    0.985515
          RI |      1.00    0.995760
    grade_01 |      1.00    0.996050
     charter |      1.00    0.999992
-------------+----------------------
    Mean VIF |      1.30
```

Figure 12: National Regression Including All State Fixed Effects for Reading

```
Linear regression                                       Number of obs = 5063770

                                                        F( 56,5063713) = 9016.60
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.1775
                                                        Root MSE      =  .52191

------------------------------------------------------------------------------
             |               Robust
   grz_state |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    z_origin |  -.2829383   .0004364  -648.37   0.000    -.2837936    -.282083
     charter |   .0073512   .0004641    15.84   0.000     .0064417    .0082608
       lunch |  -.0903302   .0005636  -160.28   0.000    -.0914348   -.0892256
         ell |  -.1916512   .0012115  -158.19   0.000    -.1940257   -.1892766
        sped |  -.2422272   .0012849  -188.51   0.000    -.2447456   -.2397088
    retained |  -.0839467   .0030309   -27.70   0.000    -.0898871   -.0780062
    re_black |  -.1314134   .0007663  -171.49   0.000    -.1329153   -.1299114
     re_hisp |  -.0539853   .0006848   -78.83   0.000    -.0553275   -.0526431
  re_asianpi |   .0835359   .0011399    73.29   0.000     .0813018    .0857701
  re_nativam |  -.1245224    .003929   -31.69   0.000    -.1322232   -.1168217
    re_multi |  -.0193101   .0020745    -9.31   0.000    -.0233761   -.0152441
       _cons |   .1499091   .0031058    48.27   0.000     .1438218    .1559965
------------------------------------------------------------------------------
```

Figure 13: VIF Test for Read Regression with All State Fixed Effects

```
Variable |       VIF       1/VIF
-------------+----------------------
      CA |     41.24    0.024247
      TX |     21.04    0.047529
      FL |     18.41    0.054327
      MI |     12.94    0.077288
**** Excluded for Space ****
re_multi |      1.03    0.970880
re_nativam |    1.02    0.981357
grade_01 |      1.00    0.996256
 charter |      1.00    0.999931
-------------+----------------------
Mean VIF |      4.15
```

Figure 14: National Regression Excluding State Fixed Effects with VIF > 10 for Reading

```
Linear regression                                      Number of obs = 5063770
                                                       F( 52,5063717) = 9305.11
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1766
                                                       Root MSE      =   .5222

------------------------------------------------------------------------------
             |               Robust
   grz_state |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    z_origin |  -.2819269   .0004289  -657.39   0.000    -.2827674   -.2810863
     charter |   .0073383   .0004643    15.80   0.000     .0064282    .0082483
       lunch |  -.0928687   .0005611  -165.51   0.000    -.0939684   -.0917689
         ell |  -.1941439   .0012052  -161.08   0.000    -.1965062   -.1917817
        sped |  -.2396231   .0012839  -186.63   0.000    -.2421396   -.2371066
    retained |  -.0865171   .0030689   -28.19   0.000    -.0925321   -.0805021
    re_black |  -.1361506   .0007524  -180.95   0.000    -.1376253   -.1346758
     re_hisp |  -.0510457   .0006623   -77.07   0.000    -.0523439   -.0497476
  re_asianpi |   .0814788   .0011326    71.94   0.000     .0792589    .0836986
  re_nativam |  -.1244079   .0039192   -31.74   0.000    -.1320895   -.1167264
    re_multi |  -.0176798   .0020738    -8.53   0.000    -.0217443   -.0136153
       _cons |   .1129899    .000954   118.44   0.000     .1111201    .1148596
------------------------------------------------------------------------------
```

Figure 15: VIF Test for Reading Regression without Four Large State Dummies

```
    Variable |       VIF       1/VIF
-------------+----------------------
   year_2009 |      2.60    0.384924
   year_2008 |      2.44    0.410622
   year_2007 |      2.30    0.435579
   year_2006 |      2.14    0.466296
       **** Excluded for Space ****
          NV |      1.01    0.988327
          RI |      1.00    0.995571
    grade_01 |      1.00    0.996282
     charter |      1.00    0.999932
-------------+----------------------
    Mean VIF |      1.31
```

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 needed 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.

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[14]. 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 6 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.

[14] 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.

Table 6: 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	5.22	7	2.20
Reading – 5 year	5.25	7	2.19
Math – 3 year	4.98	6	2.24
Math – 5 year	5.02	6	2.23

Regression Output for 16 States Comparison

Table 7: Overall Charter Effect in Reading for the 16 States

	2013 Continuing Schools		2013 New Schools		2013 All Schools	
Variable	Coefficient	SE	Coefficient	SE	Coefficient	SE
Starting score	-0.287**	(0.001)	-0.310**	(0.002)	-0.290**	(0.001)
Charter student	0.010**	(0.001)	-0.010**	(0.002)	0.007**	(0.001)
Is in Poverty	-0.093**	(0.001)	-0.101**	(0.002)	-0.095**	(0.001)
Is English Learner	-0.198**	(0.002)	-0.205**	(0.004)	-0.198**	(0.002)
Is Special Ed	-0.222**	(0.002)	-0.267**	(0.005)	-0.229**	(0.002)
Repeated grade	-0.116**	(0.005)	-0.142**	(0.011)	-0.122**	(0.004)
Black	-0.123**	(0.001)	-0.136**	(0.003)	-0.126**	(0.001)
Hispanic	-0.042**	(0.001)	-0.049**	(0.002)	-0.044**	(0.001)
Asian/Pacific Islander	0.086**	(0.002)	0.091**	(0.004)	0.087**	(0.001)
Native American	-0.130**	(0.005)	-0.150**	(0.015)	-0.131**	(0.005)
Multi-ethnic	-0.016**	(0.003)	-0.018**	(0.007)	-0.016**	(0.003)
AR	-0.064**	(0.006)	-0.069**	(0.011)	-0.070**	(0.005)
AZ	-0.028**	(0.004)	-0.002	(0.010)	-0.025**	(0.004)
CA	-0.051**	(0.004)	-0.062**	(0.009)	-0.053**	(0.004)
CO	-0.071**	(0.005)	-0.133**	(0.011)	-0.084**	(0.005)

Variable	2013 Continuing Schools		2013 New Schools		2013 All Schools	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
DC	0.060**	(0.005)	0.081**	(0.011)	0.064**	(0.004)
FL	0.001	(0.004)	-0.003	(0.010)	0.001	(0.004)
GA	-0.020**	(0.004)	-0.026*	(0.010)	-0.022**	(0.004)
IL	-0.028**	(0.005)	-0.039**	(0.010)	-0.035**	(0.004)
LA	0.012*	(0.005)	-0.088**	(0.011)	-0.010*	(0.004)
MA	-0.049**	(0.004)	-0.048**	(0.011)	-0.047**	(0.004)
MN	-0.095**	(0.005)	0.004	(0.011)	-0.083**	(0.004)
MO	-0.136**	(0.005)	-0.142**	(0.014)	-0.134**	(0.005)
NC	-0.049**	(0.004)	-0.038**	(0.011)	-0.045**	(0.004)
OH	-0.132**	(0.004)	-0.129**	(0.011)	-0.130**	(0.004)
TX	-0.062**	(0.004)	-0.001	(0.010)	-0.050**	(0.004)
year_2006	-0.002	(0.002)	-0.039**	(0.008)	-0.011**	(0.002)
year_2008	-0.004**	(0.001)	0.001	(0.003)	-0.005**	(0.001)
year_2009	0.010**	(0.001)	0.030**	(0.003)	0.011**	(0.001)
grade_01	0.549**	(0.048)	0.812**	(0.149)	0.570**	(0.046)
grade_02	0.041**	(0.007)	0.063**	(0.019)	0.044**	(0.006)
grade_03	0.016**	(0.002)	0.033**	(0.005)	0.017**	(0.002)
grade_04	-0.007**	(0.001)	-0.023**	(0.003)	-0.011**	(0.001)
grade_06	0.016**	(0.001)	0.022**	(0.003)	0.016**	(0.001)
grade_07	0.026**	(0.001)	0.029**	(0.003)	0.026**	(0.001)
grade_08	0.017**	(0.001)	0.036**	(0.003)	0.019**	(0.001)
grade_09	0.036**	(0.002)	0.051**	(0.003)	0.038**	(0.001)
grade_10	-0.030**	(0.001)	-0.035**	(0.004)	-0.031**	(0.001)
grade_11	-0.094**	(0.002)	-0.080**	(0.005)	-0.092**	(0.002)
grade_12	-1.661**	(0.012)	-1.382**	(0.027)	-1.619**	(0.011)
Constant	0.158	(0.004)	0.141**	(0.010)	0.0159**	(0.004)

Variable	2013 Continuing Schools		2013 New Schools		2013 All Schools	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Observations	2,397,932		453,312		2,851,244	
Adjusted R-squared	0.191		0.192		0.190	

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

Table 8: Overall Charter Effect in Math for the 16 States

Variable	2009 Schools since then		New Schools		Both Cohorts	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Charter student	-0.008**	(0.001)**	-0.025**	(0.002)**	-0.011**	(0.001)**
Starting score	-0.267**	(0.001)**	-0.279**	(0.001)**	-0.268**	(0.001)**
Black	-0.112**	(0.001)**	-0.127**	(0.003)**	-0.115**	(0.001)**
Hispanic	-0.025**	(0.001)**	-0.035**	(0.002)**	-0.027**	(0.001)**
Asian/Pacific Islander	0.150**	(0.002)**	0.172**	(0.004)**	0.153**	(0.002)**
Native American	-0.125**	(0.006)**	-0.118**	(0.017)**	-0.124**	(0.006)**
Multi-ethnic	-0.014**	(0.004)**	-0.025**	(0.009)**	-0.015**	(0.003)**
Is Special Ed	-0.183**	(0.002)**	-0.203**	(0.004)**	-0.186**	(0.002)**
Is English Learner	-0.083**	(0.002)**	-0.080**	(0.004)**	-0.082**	(0.002)**
Is in Poverty	-0.079**	(0.001)**	-0.068**	(0.002)**	-0.078**	(0.001)**
Repeated grade	-0.032**	(0.004)**	-0.051**	(0.008)**	-0.037**	(0.003)**
AR	-0.027**	(0.006)**	-0.072**	(0.011)**	-0.044**	(0.005)**
AZ	0.002	(0.004)	0.021*	(0.010)*	0.005	(0.004)
CA	-0.035**	(0.004)**	-0.060**	(0.010)**	-0.039**	(0.004)**
CO	0.012*	(0.005)*	-0.047**	(0.012)**	0.001	(0.005)
DC	0.121**	(0.005)**	0.081**	(0.012)**	0.114**	(0.005)**
FL	0.039**	(0.004)**	0.009	(0.010)	0.035**	(0.004)**
GA	0.012**	(0.004)**	0.016	(0.010)	0.014**	(0.004)**

Variable	2009 Schools since then		New Schools		Both Cohorts	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
IL	0.007	(0.005)	0.000	(0.010)	0.004	(0.004)
LA	0.079**	(0.005)**	-0.003	(0.011)	0.062**	(0.004)**
MA	0.011*	(0.004)*	0.033**	(0.011)**	0.015**	(0.004)**
MN	-0.080**	(0.005)**	-0.016	(0.012)	-0.072**	(0.004)**
MO	-0.129**	(0.005)**	-0.153**	(0.015)**	-0.129**	(0.005)**
NC	-0.014**	(0.004)**	-0.013	(0.011)	-0.013**	(0.004)**
OH	-0.110**	(0.004)**	-0.108**	(0.011)**	-0.108**	(0.004)**
TX	0.017**	(0.004)**	0.033**	(0.010)**	0.021**	(0.004)**
grade_01	0.569**	(0.067)**	0.600**	(0.148)**	0.574**	(0.063)**
grade_02	0.040**	(0.006)**	0.049**	(0.016)**	0.040**	(0.006)**
grade_03	-0.011**	(0.002)**	0.031**	(0.005)**	-0.004*	(0.002)*
grade_04	-0.002	(0.001)	-0.014**	(0.003)**	-0.004**	(0.001)**
grade_06	0.022**	(0.001)**	0.039**	(0.003)**	0.023**	(0.001)**
grade_07	0.052**	(0.001)**	0.064**	(0.003)**	0.053**	(0.001)**
grade_08	0.094**	(0.001)**	0.119**	(0.003)**	0.097**	(0.001)**
grade_09	-0.043**	(0.002)**	-0.012**	(0.004)**	-0.038**	(0.002)**
grade_10	-0.188**	(0.002)**	-0.123**	(0.004)**	-0.177**	(0.001)**
grade_11	-0.282**	(0.002)**	-0.192**	(0.005)**	-0.268**	(0.002)**
grade_12	-0.555**	(0.008)**	-0.559**	(0.021)**	-0.557**	(0.007)**
year_2006	-0.005**	(0.002)**	-0.009	(0.006)	-0.009**	(0.002)**
year_2008	-0.001	(0.001)	-0.006*	(0.003)*	-0.003**	(0.001)**
year_2009	0.012**	(0.001)**	0.028**	(0.003)**	0.012**	(0.001)**
Constant	0.079**	(0.004)**	0.058**	(0.010)**	0.078**	(0.004)**
Observations	2,258,392		438,726		2,697,118	
Adjusted R-squared	0.162		0.168		0.163	

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

Regression Output for 27 States

Table 9 National (27 State) Regression Output, Overall Models

Variable	Reading		Math	
	Coefficient	SE	Coefficient	SE
Charter Student	0.010**	(0.002)	-0.005	(0.004)
Starting Score	-0.277**	(0.002)	-0.261**	(0.003)
Black	-0.121**	(0.004)	-0.114**	(0.004)
Hispanic	-0.046**	(0.003)	-0.033**	(0.004)
Asian or Pacific Islander	0.085**	(0.005)	0.151**	(0.007)
Native American	-0.119**	(0.009)	-0.113**	(0.011)
Multi-Ethnic	-0.018**	(0.004)	-0.022**	(0.005)
Is Special Ed	-0.240**	(0.004)	-0.190**	(0.003)
Is English Learner	-0.184**	(0.005)	-0.075**	(0.004)
Is In Poverty	-0.090**	(0.002)	-0.093**	(0.001)
Repeated Grade	-0.069**	(0.019)	0.004	(0.014)
AR	-0.063**	(0.019)	-0.042	(0.024)
AZ	-0.030**	(0.019)	-0.003	(0.024)
CA	-0.064**	(0.013)	-0.052*	(0.022)
CO	-0.071**	(0.015)	-0.034	(0.024)
DC	0.053**	(0.017)	0.105**	(0.026)
FL	-0.009	(0.013)	0.026	(0.022)
GA	-0.033*	(0.015)	-0.001	(0.025)
IN	-0.077**	(0.023)	-0.049	(0.026)
IL	-0.021	(0.014)	0.016	(0.022)
LA	-0.020	(0.020)	0.051**	(0.026)
MA	-0.051**	(0.016)	0.006	(0.024)
MI	-0.096**	(0.013)	-0.063**	(0.022)
MN	-0.086**	(0.016)	-0.076**	(0.023)

Variable	Reading		Math	
	Coefficient	SE	Coefficient	SE
MO	-0.137**	(0.019)	-0.135**	(0.027)
NC	-0.069**	(0.015)	-0.031	(0.023)
NJ	-0.051**	(0.020)	-0.015	(0.023)
NV	-0.083**	(0.036)	0.028	(0.056)
NX	-0.110**	(0.016)	-0.068**	(0.025)
NY	0.006	(0.014)	0.096**	(0.024)
OH	-0.130**	(0.014)	-0.113**	(0.022)
OR	-0.077**	(0.015)	-0.056*	(0.023)
PA	-0.123**	(0.016)	-0.107**	(0.026)
RI	-0.074**	(0.027)	0.004	(0.025)
TN	-0.142**	(0.019)	-0.060	(0.035)
TX	-0.063**	(0.014)	0.009	(0.022)
UT	-0.116**	(0.015)	-0.068**	(0.025)
grade_01	0.568**	(0.061)	0.572**	(0.074)
grade_02	0.073**	(0.017)	0.066**	(0.024)
grade_03	0.035**	(0.005)	0.025**	(0.007)
grade_04	0.018**	(0.003)	0.015**	(0.004)
grade_06	0.016**	(0.003)	0.025**	(0.004)
grade_07	0.021**	(0.003)	0.041**	(0.004)
grade_08	0.002	(0.003)	0.070**	(0.008)
grade_09	0.045**	(0.007)	-0.033**	(0.011)
grade_10	-0.025**	(0.006)	-0.170**	(0.010)
grade_11	-0.068**	(0.010)	-0.249**	(0.011)
grade_12	-1.402**	(0.083)	-0.429**	(0.030)
year_2008	-0.005**	(0.002)	-0.005*	(0.002)
year_2009	0.003	(0.003)	0.006*	(0.003)
period_1	-0.058**	(0.002)	-0.050**	(0.004)

Variable	Reading		Math	
	Coefficient	SE	Coefficient	SE
period_2	-0.018**	(0.002)	-0.013**	(0.004)
period_4	0.004	(0.002)	0.010**	(0.004)
period_5	0.019	(0.004)	0.014**	(0.007)
Constant	0.189**	(0.013)	0.113**	(0.022)
Observations	3,483,732		3,346,524	
Adjusted R-squared	0.168		0.155	

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

Table 10: National (27 State) Regression Output, Sub-Population Models

	Reading		Math	
Variable Label	Coefficient	SE	Coefficient	SE
Starting score	-0.277**	(0.003)	-0.262**	(0.002)
Charter Black	-0.123**	(0.005)	-0.138**	(0.007)
TPS Black	-0.139**	(0.003)	-0.155**	(0.004)
Charter Hispanic	-0.059**	(0.005)	-0.071**	(0.007)
TPS Hispanic	-0.055**	(0.003)	-0.059**	(0.004)
Charter Asian or Pacific Islander	0.070**	(0.008)	0.102**	(0.011)
TPS Asian or Pacific Islander	0.081**	(0.004)	0.137**	(0.007)
Charter Native American	-0.145**	(0.013)	-0.184**	(0.019)
TPS Native American	-0.114**	(0.010)	-0.108**	(0.010)
Charter White	-0.021**	(0.004)	-0.065**	(0.007)
Charter – Special Ed	-0.236**	(0.005)	-0.183**	(0.004)
TPS – Special Ed	-0.245**	(0.004)	-0.199**	(0.003)
Charter – English Learner	-0.161**	(0.006)	-0.052**	(0.006)
TPS – English Learner	-0.208**	(0.005)	-0.098**	(0.004)
Charter – in Poverty	-0.081**	(0.003)	-0.063**	(0.005)
TPS – in Poverty	-0.099**	(0.002)	-0.091**	(0.003)
Charter – Repeated Grade	-0.046	(0.024)	0.006	(0.016)
TPS – Repeated Grade	-0.093**	(0.017)	0.001	(0.015)
AR	-0.063**	(0.017)	-0.042**	(0.022)
AZ	-0.030*	(0.013)	-0.003	(0.021)
CA	-0.064**	(0.012)	-0.052**	(0.019)
CO	-0.071**	(0.014)	-0.034**	(0.021)
DC	0.053**	(0.016)	0.105**	(0.024)
FL	-0.009	(0.012)	0.026	(0.019)
GA	-0.033*	(0.014)	-0.001	(0.022)

Variable Label	Reading		Math	
	Coefficient	SE	Coefficient	SE
IN	-0.077**	(0.019)	-0.049*	(0.022)
IL	-0.021	(0.014)	0.016	(0.020)
LA	-0.030	(0.017)	0.051*	(0.022)
MA	-0.051**	(0.015)	0.006	(0.022)
MI	-0.096**	(0.013)	-0.063**	(0.019)
MN	-0.086**	(0.014)	-0.076**	(0.021)
MO	-0.137**	(0.017)	-0.135**	(0.024)
NC	-0.069**	(0.013)	-0.031	(0.020)
NJ	-0.051**	(0.018)	-0.015	(0.024)
NV	-0.084**	(0.040)	0.028	(0.049)
NX	-0.111**	(0.016)	-0.068**	(0.025)
NY	0.006	(0.014)	0.096**	(0.021)
OH	-0.131**	(0.014)	-0.113**	(0.021)
OR	-0.077**	(0.014)	-0.056**	(0.020)
PA	-0.123**	(0.015)	-0.107**	(0.023)
RI	-0.074**	(0.027)	0.004	(0.029)
TN	-0.142**	(0.020)	-0.060	(0.035)
TX	-0.063**	(0.013)	0.009	(0.020)
UT	-0.116**	(0.014)	-0.068**	(0.022)
grade_01	0.568**	(0.057)	0.572**	(0.067)
grade_02	0.073**	(0.073)	0.066**	(0.020)
grade_03	0.035**	(0.004)	0.025**	(0.006)
grade_04	0.018**	(0.003)	0.015**	(0.004)
grade_06	0.016**	(0.003)	0.025**	(0.004)
grade_07	0.021**	(0.003)	0.041**	(0.004)
grade_08	0.002	(0.003)	0.070**	(0.007)
grade_09	0.045**	(0.006)	-0.033**	(0.009)

Variable Label	Reading		Math	
	Coefficient	SE	Coefficient	SE
grade_10	-0.025**	(0.005)	-0.170**	(0.008)
grade_11	-0.068**	(0.008)	-0.249**	(0.009)
grade_12	-1.401**	(0.060)	-0.429**	(0.022)
year_2008	-0.005**	(0.002)	-0.005*	(0.002)
year_2009	0.003*	(0.002)	0.006*	(0.002)
period_2	0.041**	(0.002)	0.037**	(0.003)
period_3	0.058**	(0.002)	0.051**	(0.004)
period_4	0.062**	(0.003)	0.060**	(0.005)
period_5	0.077**	(0.005)	0.064**	(0.007)
Constant	0.146**	(0.012)	0.093**	(0.019)
Observations	3,483,748		3,346,530	
Adjusted R-squared	0.169		0.157	

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

Response to The Center for Education Reform's Criticisms

Table 11: A Point by Point Correction of The Center for Education Reform's Criticisms

Center for Education Reform Critique	CREDO Comments
The Virtual Twin method is not "the gold standard" and is therefore not capable of providing accurate estimates of charter impact on student growth.	As is discussed in the "Model Selection" section of the technical appendix, independent analyses of the VCR method have shown that it provides estimates of charter effects at least as close to the "gold standard" of charter research (lottery studies) as any other quasi-experimental approach. In fact, the effects found using the VCR method were not significantly different than those found using a lottery approach *on the same exact students*. Combined with the external validity possible by including a higher percentage of charter students in our data in the analysis (not possible in lottery analyses), CREDO has confidence that the VCR method is capable of providing accurate estimates of charter impact.
The Virtual twin method relies on fake children for gauging learning gains.	This statement reflects a misunderstanding of how Virtual Control Records are created. The virtual twins' achievement values are composites of real students. All the values other than test scores must match perfectly with each charter student record. Test scores often match exactly but can vary by 0.1 standard deviations (An independent analysis showed that this variation does not bias estimates of charter effect). The VCR outcomes are simply averages of up to 7 students rather than of a single student. This provides a more stable comparison than matching to a single TPS student and according to independent analyses does not systematically bias estimates of charter effect either up or down.
CREDO's analysis is not a "gold standard" randomized lottery trial.	Randomized controlled trials (RCT) are considered the "gold standard" in social science research. However, there are a few caveats necessary to conduct a RCT. 1. **The lottery must be random.** This is often not true in charter schools, as many schools permit preferences to siblings of current students, children of school founders or staff, or residential preferences for students who live near the school (See Betts, J. and Hill, P., 2006) for a summary of potential challenges to the internal validity of

Center for Education Reform Critique	CREDO Comments
	RCTs). 2. **There must be sufficient numbers of students participating in the lottery.** In other words, there have to be large numbers of students who do not get selected into the charter school. While many charter schools have waiting lists, most charter schools do not have large enough waiting lists for a RCT. 3. **The charter schools that meet conditions 1 & 2 above must be representative of all charter schools.** Violating condition 3 creates major problems in conducting a valid **national** charter study using RCT for several reasons. a. Charter schools which have a long term reputation for quality may be more likely to hold a lottery than weaker or newer charter schools. b. Charter schools located near particularly low quality traditional public schools may be more likely to hold lotteries than charters located near higher performing TPS. c. Charter schools in areas with fewer choice options may be more likely to have lotteries than charter schools located in areas with a higher number of choice options. 4. **RCTs have strong internal validity but weaker external validity.** While RCTs are the gold standard for estimating the effect of a single treatment (e.g. the effect of attending a specific charter school), any of the violations listed in 3 above could damage the ability to generalize results to other charter schools. CREDO's matching method has much greater external validity because it is not limited to charter schools with random lotteries and sufficiently large waiting lists. The charter schools and students in CREDO's data set look much more like the national charter sector than those

Center for Education Reform Critique	CREDO Comments
	eligible to be included in a RCT. In addition, a recent meta-analysis of the charter school literature found that, "as long as baseline test scores are controlled for, the specific method of analysis employed will not severely impact conclusions." (Betts, J et al., 2011) In this light, RCTs and quasi-experimental methods should be considered complements, not substitutes.
CREDO used NAEP scores to create the virtual twins rather than individual test scores.	It is unclear how this notion arose. CREDO has never used NAEP scores to create VCRs. The VCRs are created using individual student level data acquired directly from each state's department of education.
It is statistically impossible to come even close to a "virtual twin" for 20 to 25 percent of charter school students.	This is inaccurate. The match rate in the 2013 national report is 86% in reading and 84% in math. This gives the VCR study a far higher inclusion rate than either a lottery or fixed effects estimation method could provide.
CREDO's report does not take into account the higher percentage of charter elementary and middle schools, leading to inaccurately weighted aggregate data.	Given that the sample used in the CREDO study pulls from all school levels equally, this assertion is groundless. The data set used in CREDO's analysis reflects the proportion of schools at each level in the national charter sector.
CREDO's overall effect is skewed because of the small number of high schools in the charter sector.	For this critique to be valid, CREDO would have had to report only a simple average effect by school type. The overall effects reported in CREDO's national reports are aggregated student-level results.
Most charter schools are not classified as they are "multi-level" schools.	This critique can be dismissed with publicly available data. Data from the National Alliance for Public Charter Schools shows that in 2010-11 44% of charter schools are elementary schools, 20% are high schools, 10% are middle schools and 26% are a type of multi-level configuration.
Long-term studies demonstrate strong growth for students who stay in charter schools.	To justify this statement, the Center for Education Reform chooses particular results from the broad array of charter school research. It is generally advisable when attempting to identify the "real" effect of a heterogeneous impact (like charter school impacts) that one look at the range of high quality studies conducted on the subject, rather than a subset of the literature.
CREDO's study has been discredited by Caroline Hoxby for not meeting the	The "statistical mistake" identified by Dr. Hoxby has been tested and shown irrelevant to the validity of CREDO's

Center for Education Reform Critique	CREDO Comments
"gold standard" of being an experimental design.	estimation of charter effect (See "Model Selection" and "Section G." of the technical appendix). In addition, multiple independent analyses have found that the VCR method produced results which were not significantly different from the results of experimental methods.
CREDO's analysis does not account for the great variances in charter laws from state to state or how those laws may differ from paper to practice.	CREDO has released a large number of state-level studies. These state-level studies are available on the same webpage as the national study. CREDO's standard model also includes state level fixed effects (which control for time invariant state level factors, such as policy environments). Models were also run with state and year fixed effects interacted to capture the potential impact of changes in state policy environment by year, and these were found to have no impact on estimates of charter growth.
Not an apples to apples comparison.	The use of feeder lists to create matches means that charter students are only matched with students from TPS that their charter schools' students previously attended. Further, matching on prior year test scores ensures that the educational preparation of charter students and their matches are the same on average.
Ignored differences between state test rigor and data.	Test scores were standardized by state before being pooled for a national analysis (see "Data" section of technical appendix). CREDO's baseline model also included controls for different environments in each state.
Use of FRL eligibility data may not properly represent the range of student poverty.	The vast majority of studies of academic performance use this data to identify students in poverty. Matching on each student's prior year test score ensures that even if students are matched on an imperfect proxy for poverty, charter students are being matched only with other students based on the total educational preparation of each student, not on any single variable that may impact achievement and growth.

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