NEPC

NATIONAL EDUCATION POLICY CENTER

Review of *National Charter School Study 2013*

Reviewed By

Andrew Maul and Abby McClelland

University of Colorado Boulder

July 2013

Summary of Review

The Center for Research on Education Outcomes (CREDO) at Stanford University analyzed differences in student performance at charter schools and traditional public schools across 27 states and New York City. The study finds a small positive effect of being in a charter school on reading scores and no impact on math scores; it presents these results as showing a relative improvement in average charter school quality since CREDO's 2009 study. However, there are significant reasons for caution in interpreting the results. Some concerns are technical: the statistical technique used to compare charter students with "virtual twins" in traditional public schools remains insufficiently justified, and may not adequately control for "selection effects" (i.e., families selecting a charter school may be very different from those who do not). The estimation of "growth" (expressed in "days of learning") is also insufficiently justified, and the regression models fail to correct for two important violations of statistical assumptions. However, even setting aside all concerns with the analytic methods, the study overall shows that less than one hundredth of one percent of the variation in test performance is explainable by charter school enrollment. With a very large sample size, nearly any effect will be *statistically* significant, but in *practical* terms these effects are so small as to be regarded, without hyperbole, as trivial.

Kevin Welner
Project Director

William Mathis
Managing Director

Erik Gunn
Managing Editor

National Education Policy Center

School of Education, University of Colorado
Boulder, CO 80309-0249
Telephone: (802) 383-0058

Email: NEPC@colorado.edu
http://nepc.colorado.edu

Publishing Director: Alex Molnar

This is one of a series of Think Twice think tank reviews made possible in part by funding from the Great Lakes Center for Education Research and Practice. It is also available at http://greatlakescenter.org.

Review of *National Charter School Study 2013*

Andrew Maul and Abby McClelland, University of Colorado Boulder

I. Introduction

Since 2009, the Center for Research on Education Outcomes (CREDO) at Stanford University has produced a series of reports on the performance of charter schools relative to traditional public schools (TPSs). These reports seek to inform ongoing conversations among policymakers and researchers regarding whether charter schools are likely to generate better outcomes than TPSs overall. The reports also explore whether these effects might be especially pronounced for members of particular subgroups, such as students from minority racial/ethnic backgrounds and from less socioeconomically advantaged backgrounds.

CREDO's latest study, *National Charter School Study 2013*,[1] employs a methodological approach highly similar to their previous reports. The study concludes that, overall, charter schools perform better now than they had in 2009 in terms of students' scores on mathematics and reading tests, to the point they now perform on par with traditional public schools (students in charter schools were estimated to score approximately 0.01 standard deviations higher on reading tests and 0.005 standard deviations lower on math tests than their peers in TPSs).

II. Findings and Conclusions of the Report

Based on data on 1,532,506 charter students and a matched group of students in traditional public schools furnished by the departments of education from the 27 states examined in this study (which are stated, collectively, to contain 95% of the nation's charter students), the report presents a large number of conclusions, including the following:

- On average, it was estimated that students in charter schools in the 16 states studied in 2009 fare better on academic tests relative to their peers in traditional public schools in 2013 than they did in 2009. On reading tests, charter students were estimated to score 0.01 standard deviations lower than their TPS peers in 2009, but such students are estimated to score 0.01 standard deviations higher in

2013; on mathematics tests, charter students were estimated to score 0.03 standard deviations lower in 2009, but such students are estimated to score only 0.01 standard deviations lower in 2013. This is taken as evidence that charter schools are improving relative to TPSs.

- At least two explanations are available for the apparent improvement of charter schools relative to traditional public schools. The first is the reported overall decline in performance at TPSs. The second is the closing of lower-performing charter schools (approximately 8% of the 2009 sample).
- When taken together across the 27 states, students in charter schools were estimated to score approximately 0.01 standard deviations higher on reading tests and 0.005 standard deviations lower on math tests than their peers in TPSs.
- There were significant state-to-state variations in the estimated differences between charter and TPSs.
- The apparent advantage of charter school enrollment was estimated to be slightly greater on average for students in poverty and Hispanic English Language Learners. Conversely, White students, Asian students, and non-ELL Hispanic students appeared to fare slightly worse in charter schools than their peers in TPSs.
- The advantage of being enrolled in a charter school appeared to increase as a function of the number of years a student was enrolled in a charter school.
- When considering school averages rather than individual students, 25% of charter schools were estimated to have greater average growth than comparable TPSs in reading, and 29% were estimated to have greater average growth in math. Nineteen percent of charter schools were estimated to have lower average growth in reading, and 31% were estimated to have lower average growth in math.

III. The Report's Rationale for Its Findings and Conclusions

Like previous CREDO studies, this is an empirical study. The conclusions are based primarily on analyses of datasets furnished by the 27 state departments of education, which are collectively stated to include observations from 1,532,506 charter school students along with a matched group of comparison students from traditional public schools. These results are also compared with the findings from the 2009 study.

Overall, the study concludes that "while much ground remains to be covered, charter schools in the 27 states are outperforming their TPS peer schools in greater numbers than in 2009" (p. 85). Although no explicit policy recommendations are stated in the report, a variety of "implications" are explored, the most prominent of which is that the improvement of the charter school sector relies on the setting of high standards and the selective closure of under-performing schools.

The data-collection and analytic methods are described to some extent in the main report, and further detail is given in a technical appendix. The primary rationales for the study's conclusions are based on the results of a series of regression models that attempt to compare students in charter schools with students in traditional public schools who are matched on a set of seven background characteristics. These analytic methods will be discussed further in section V, below.

IV. The Report's Use of Research Literature

As with previous state-level CREDO reports on charter school data, the contents of the report focus on their findings. The report does not contain a literature review and contains minimal reference to other evidence, save CREDO's earlier studies.

V. Review of the Report's Methods

Miron and Applegate[2] and Maul[3], in their earlier reviews, have called attention to a variety of technical and conceptual concerns with the methods employed by the CREDO charter school studies. CREDO researchers have not altered their methodology in light of those concerns; thus, many of the comments we make here overlap with previously raised issues. Here we comment on four technical issues: (a) the approach used to match charter students with "virtual twins" for comparison, (b) the modeling of the multilevel structure of the data and measurement error, (c) the estimation of growth, and (d) unexplained and apparently arbitrary analytic choices.

Concerns about the Matching Procedure

Defending a causal inference in the absence of a controlled experimental design requires an argument that observational data can be used to provide an estimate of the counterfactual, or what would have happened to charter school students had they attended a traditional public school. CREDO's argument depends on the construction of a "Virtual Control Record" (VCR) for each student in a charter school, obtained by averaging together up to seven students in "feeder" public schools (i.e., those schools whose students transfer to charters) with the same gender, ethnicity, English proficiency status, eligibility for subsidized meals, special education status, grade level, and a similar score from a prior year's standardized test (within a tenth of a standard deviation) as the specified charter student.

It is unclear why CREDO decided to develop and use a home-grown matching technique (the VCR method) given the availability of propensity-based matching techniques,[4] which would seem to be superior in several respects. The VCR technique requires exact matches, whereas propensity-based methods do not. This means that with propensity matching,

arbitrary cutoffs for continuous covariates (e.g., matching to twins with a prior year test score within 0.1 standard deviations) are unnecessary.

Even more troubling, the VCR technique found a match for only 85% of charter students. There is evidence that the excluded 15% are in fact significantly different from the included

> *The bottom line appears to be that, once again, it has been found that, in aggregate, charter schools are basically indistinguishable from traditional public schools in terms of their impact on academic test performance.*

students in that their average score is 0.43 standard deviations lower than the average of the included students; additionally, members of some demographic subgroups such as English Language Learners were much less likely to have virtual matches.

No clear explanation[5] is given for these worrying differences. A propensity-based method would probably have allowed inclusion of far more than 85% of the charter sample. Furthermore, although it is shown in the technical appendix that starting scores of charter students are similar to their TPS counterparts, no information is reported regarding the success of the matching procedure in eliminating selection bias as measured by any of the other matching variables.

The larger issue with the use of any matching-based technique is that it depends on the premise that the matching variables account for all relevant differences between students; that is, once students are matched on the aforementioned seven variables,[6] the only remaining meaningful difference between students is their school type. Thus, for example, one must believe that there are no remaining systematic differences in the extent to which parents are engaged with their children (despite the fact that parents of charter school students are necessarily sufficiently engaged with their children's education to actively select a charter school), that eligibility for subsidized meals is a sufficient proxy for poverty when taken together with the other background characteristics,[7] and so forth.

CREDO cites two papers in defense of the use of VCRs. The first of these[8] compares a study based on lottery analysis (which is argued to be closer to a "gold standard" experimental situation) with four forms of non-experimental estimation of charter school effects, including the VCR approach, and concludes that the relatively small differences in the effects found by the VCR and lottery methods were not statistically significantly different from one another (given the sample size of their study). Interestingly, in that context the magnitude of effect sizes estimated by the VCR technique were 0.02 standard deviations in math and 0.04 in reading, compared with -0.01 in math and 0.00 in reading for the lottery study. Although we agree that the differences of 0.03 and 0.04 between the VCR and lottery analyses are small, they are considerably larger than the 0.02 standard deviation differences between the 2009 and 2013 CREDO studies, which form the basis for the purportedly newsworthy conclusions of the new study!

The second cited paper compares a lottery study in New York City with a CREDO study in the same city the following year.[9] This is potentially more convincing, as this CREDO study, like the present study, compares charter students with traditional public students who likely did not apply for charter school attendance, and the CREDO approach does produce a similar result to the lottery-based study. However, once again, estimated coefficients differ by as much as 0.03 standard deviations. Also, the report notes that the situation in New York City was unusual,[10] and it is therefore not clear that these findings can be generalized with confidence.

Thus, one must take as an article of faith that the seven matching variables used in the construction of VCRs truly captured all important differences between charter and TPS students; readers finding this implausible will be unconvinced that these methods can provide accurate estimates of causal effects.

Concerns with Two Violations of Statistical Assumptions

Two essential criteria for the use of regression models are (a) that observations are independent, and (b) that all variables are measured without error. Both of these assumptions are violated in the data analyzed in this report, and it does not appear that the analyses were adjusted appropriately.

With respect to the first concern (independence of observations), the data analyzed are hierarchical in the sense that students are nested within classrooms and schools, and it is likely that observations are not conditionally independent. That is, there is likely to be considerable within-school shared variance, since individual records used in the study would be for students sharing the same school, and often the same teacher and classroom. Multilevel modeling seems like it would have been the natural choice for such a data structure. In the report's technical appendix, a comparison is given of the estimation of overall main effects using "robust" versus "clustered" standard errors,[11] and it is concluded that "results… are found to be comparable, making comparisons of effect sizes between the methods valid" (Technical Appendix, p. 14).[12] However, this statement apparently ignores the fact that one of the two main effects changes from being statistically significant when using the "robust" errors to non-significant when using the "clustered" option (Technical Appendix, p. 22). Thus these results show the opposite of what is claimed: they show that it very much *does* matter whether the clustered nature of the data is taken into account, which is worrying given that no evidence can be found in the report or technical appendix that this issue was addressed beyond this comparison.

Additionally, the approach used in the study demands the exclusion of schools with a lower number of tested students. The researchers report that, in 2010-2011, about 7% of the charter schools (248 out of 3670 for reading, and 266 out of 3654 for math) were excluded from analyses on the basis of having an insufficient number of tested students[13] to "calculate a representative school-wide average growth rate" (p. 55). This exclusion would not be necessary if error in the measurement of school means was explicitly modeled (as would occur in the aforementioned multilevel modeling approach). This could be

problematic if the excluded schools were systematically different from others, which seems plausible given that they are more likely to be new—and as is stated elsewhere in the report, newer charter schools fare worse on average compared with traditional public schools.

With respect to the second concern, error-free measurement, academic tests invariably contain some amount of measurement error. In itself this is not necessarily problematic as long as this error is modeled properly. In the technical appendix, it is acknowledged that the level of error in tests varies depending on state, grade, and a student's place in the distribution (e.g., a high-scoring, low-scoring, or middle-scoring student). A comparison is given of the estimation of overall main effects using (a) an "errors-in-variables" regression,[14] which explicitly models measurement error, and (b) two ordinary regression models, neither of which model measurement error, but which utilize each of the two forms of robust standard errors described earlier. This comparison is bizarre, as robust standard errors and the errors-in-variables approach are designed to deal with two completely different kinds of error—misestimation of standard errors due to violations of model assumptions in the former case, and measurement error in the latter case. Furthermore, they are perfectly compatible, in that both techniques can be used at once—which, given that both corrections appear to be warranted for different reasons, raises the question of why they were not both applied in any model.

Concerns with the Estimation of Growth

As with previous reports, findings are described in terms of "growth," estimated via average year-to-year gains on state standardized tests expressed in standard deviation units. These are translated into "days of learning" via a procedure that is never explained or discussed.[15]

The expression of differences in test scores in terms of "days of learning" requires accepting substantial untested assumptions about the nature of the student attributes measured by the state tests. This is a controversial topic in the psychometric literature, and while the report acknowledges that "the days of learning are only an estimate and should be used as general guide rather than as empirical values" (p.13), it nevertheless uses "days of learning" on the axes of all graphs and often expresses findings only in terms of "days of learning," rather than the accepted convention of standard deviation units. Without a clear (or indeed any) rationale for this choice, the "days of learning" metric cannot be regarded as credible.

Furthermore, as Miron and Applegate[16] noted, making inferences to longitudinal growth of individual students' levels of achievement also leans on other (unstated) assumptions, most notably that the group of students used to construct the Virtual Control Records is itself stable (i.e., that the VCR is constructed using essentially the same students over time). Given that the researchers had access to individual student records, changes at the level of the individual could have been modeled directly using a multilevel framework; it is unclear why this was not done.

Arbitrary and Unexplained Analytic Choices

Throughout the report, the authors mention seemingly arbitrary choices made during the analysis, and they provide no explanation or discussion of the rationale for those choices. While these choices may or may not affect the final results, they would have benefited from explicit discussion. Examples of these issues include: (a) using a lookback period of five years to examine student-level performance in the 27 states, but only a two-year lookback to examine school-level performance; (b) examining only 27 of the 43 states that authorize charter schools; (c) using fourth-grade NAEP scores as a measure of overall state achievement in grades K-12; and, perhaps most troublingly, (d) eschewing any form of test-equating across the 27 states and instead assuming that all scores can be standardized and projected onto a common metric.[17]

VI. Review of the Validity of the Findings and Conclusions

This review has noted a number of reasons for concern regarding the methodology employed in CREDO's *National Charter School Study 2013*. However, even setting aside all of these concerns, the actual magnitude of each of the effects reported in this study is extremely small. The very large sample size guarantees that nearly any predictor will be statistically significant; however, in practical terms, the differences are trivial. The most important results of the study (between the 2009 overall results and the 2013 results) are differences of 0.01 or 0.02 standard deviation units, and even the largest effect sizes reported (e.g., the estimated effect of charter schools for Hispanic English Language Learners) are on the order of 0.07 standard deviations.[18]

To put these effect sizes in context, a difference of 0.07 standard deviations between two groups means that just over one tenth of one percent (0.0012) of the variation in test scores can be attributed to whether a student is in a charter school or a traditional public school. A difference of 0.01 standard deviations indicates that a quarter of a hundredth of a percent (0.000025) of the variation can be explained. As another point of reference, Hanushek has described an effect size of 0.20 standard deviations for Tennessee's class size reform as "relatively small" considering the nature of the intervention.[19] To give a different example, a student correctly answering a single additional question (out of 54) on the SAT Math test would boost her standardized score by anywhere from 0.05 standard deviations to more than 0.30 standard deviations depending on her place in the distribution. In standard research contexts, the effect sizes in this study are so close to zero as to be regarded as effectively zero. Thus the bottom line appears to be that, once again, it has been found that, in aggregate, charter schools are basically indistinguishable from traditional public schools in terms of their impact on academic test performance.

When one also considers the methodological concerns noted above—and notes that, given the small effect sizes, even a minor methodological issue could play a decisive role—it seems clear that advocacy claims regarding the results of this study must be interpreted with extreme caution.

VII. Usefulness of the Report for Guidance of Policy and Practice

Any given study of charter schools will have strengths and weaknesses. The size and comprehensiveness of the dataset analyzed make this report an interesting contribution to the charter school research base. However, this review has noted reasons for caution when making inferences to a true causal effect of charter schools. As such, it is advised that the findings of this report not be regarded as definitive evidence of the increasing effectiveness of charter schools since 2009. In effect, this study's results are consistent with past research that suggests that charter schools are essentially indistinguishable from traditional public schools in terms of their effects on academic performance.

Notes and References

1 Center for Research on Education Outcomes (CREDO) (2013, June). *National Charter School Study*. Palo Alto: CREDO, Stanford University. Retrieved July 10, 2013, from http://credo.stanford.edu/research-reports.html.

2 Miron, G. & Applegate, B. (2009). *Review of "Multiple choice: Charter school performance in 16 states."* Boulder, CO: National Education Policy Center. Retrieved July 10, 2013, from http://nepc.colorado.edu/thinktank/review-multiple-choice/.

3 Maul, A. (2013). *Review of "Charter School Performance in Michigan."* Boulder, CO: National Education Policy Center. Retrieved July 10, 2013, from http://nepc.colorado.edu/thinktank/review-charter-performance-michigan.

4 Propensity-based score matching is an increasingly common way of attempting to reduce bias in the estimation of causal effects from observational data due to the confounding influence of variables that predict whether or not a student receives a treatment. Such techniques predict the probability of treatment based on a set of conditioning variables, which can be either continuous or categorical, and then match subjects in the two groups based on similarity in this probability; thus exact matches are not required.

5 On p.14 of the technical appendix it is stated that "this is because students at the very low and high end of the test score distribution have more trouble finding matches in traditional public schools." However, this does not explain why the omitted students score nearly half a standard deviation *lower*, on average, than the included group: if there were matching failures on both ends of the distribution, one would expect these differences to cancel out, or nearly so.

6 Studies using propensity-based methods frequently use very large numbers (e.g., 70 or greater) of variables to match students, and even then there is debate concerning whether the matches can be thought of as true counterfactuals.

7 For example, consider the difference between free and reduced-price lunch (which are lumped together in the present analysis). Children qualify for free lunch if their families are below 130% of poverty level. They qualify for reduced-price lunch if their families are below the 185% of poverty level. The present analysis must assume either that these two categories are equally distributed across charter and TPS students, or that the difference does not matter. See pp. 8-9 of:

Baker, B.D. & Ferris, R. (2011). *Adding Up the Spending: Fiscal Disparities and Philanthropy among New York City Charter Schools*. Boulder, CO: National Education Policy Center. Retrieved July 10, 2013, from http://nepc.colorado.edu/publication/NYC-charter-disparities/.

8 Forston, K. & 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. Washington, DC: U.S. Department of Education.

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

10 The study's authors note that the effects found for New York City charter schools "are quite robust and large relative to the charter school effects in most other locations" (p. 51).

11 Though it is never explicitly stated what "clustered" standard errors are, it can be surmised that this refers to the "vce(cluster id)" option in Stata (as opposed to the "robust" option). In fact these are both robust standard error methods; they differ only in that the former effectively corrects for within-cluster variance whereas the latter does not. (One must assume that the "clusters" in this case are *schools;* this is never stated.) Conceptually, there is no obvious reason why one would prefer the latter to the former, given the clustered nature of the data.

12 It is never stated which of these corrections, if either, was used in the analyses presented in the main document.

13 It is stated that the criteria were that the school have at least 60 matched charter students over a two-year period, or at least 30 matched students in the case of charter schools that have only one year of data. The justification for these particular cutoffs is not explained.

14 The details of the implementation of this approach are not given.

15 It is unclear what denominator is used for the "days of learning" metric. If one assumes that the average difference in test scores between two academic years is used as the determination of a "year of growth," is this number simply divided by 365? If so, resulting numbers will be significantly larger than if the denominator is instead the number of school days in a year.

16 Miron, G. & Applegate, B. (2009). *Review of "Multiple choice: Charter school performance in 16 states."* Boulder, CO: National Education Policy Center. Retrieved July 10, 2013, from http://nepc.colorado.edu/thinktank/review-multiple-choice/.

17 For further discussion of the potential problems of standardizing tests from multiple states without engaging in test equating, see p.6 of:

Miron, G. & Applegate, B. (2009). *Review of "Multiple choice: Charter school performance in 16 states."* Boulder, CO: National Education Policy Center. Retrieved July 10, 2013, from http://nepc.colorado.edu/thinktank/review-multiple-choice/.

18 It was noted earlier that ELL charter students with lower scores appear to have been disproportionally excluded from the study due to matching failure; it could thus be surmised that these (relatively) larger effect sizes are especially vulnerable to the methodological concerns discussed previously.

19 See p. 56 in:

Hanushek, E. A (2002). Evidence, politics, and the class size debate. In L. Mishel & R. Rothstein (eds.). *The Class Size Debate*, Washington, DC: Economic Policy Institute, 37-65. Retrieved July 10, 2013, from http://borrellid64.com/images/The_Class_Size_Debate.pdf/.

DOCUMENT REVIEWED:	**National Charter School Study 2013**
AUTHOR:	Center for Research on Education Outcomes (CREDO)
PUBLISHER:	CREDO
DOCUMENT RELEASE DATE :	June 2013
REVIEW DATE:	July 16, 2013
REVIEWERS:	Andrew Maul & Abby McClelland, University of Colorado Boulder
E-MAIL ADDRESS:	andrew.maul@colorado.edu
PHONE NUMBER:	(303) 492-7653

SUGGESTED CITATION:

Maul, A. & McClelland, A. (2013). *Review of "National Charter School Study 2013."* Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/review-credo-2013/.