NEPC

NATIONAL EDUCATION POLICY CENTER

Review of *Urban Charter School Study 2015*

Reviewed By

Andrew Maul

University of California-Santa Barbara

April 2015

Summary of Review

Following up on a previous study, researchers sought to investigate whether the effect on reading and math scores of being in a charter school was different in urban areas compared with other areas and to explore what might contribute to such differences. Overall, the study finds a small positive effect of being in a charter school on both math and reading scores and finds that this effect is slightly stronger in urban environments. There are significant reasons to exercise caution, however. The study's "virtual twin" technique is insufficiently documented, and it remains unclear and puzzling why the researchers use this approach rather than the more accepted approach of propensity score matching. Consequently, the study may not adequately control for the possibility that families selecting a charter school may be very different from those who do not. Other choices in the analysis and reporting, such as the apparent systematic exclusion of many lower-scoring students from the analyses, the estimation of growth, and the use of "days of learning" as a metric, are also insufficiently justified. Even setting aside such concerns over analytic methods, the actual effect sizes reported are very small, explaining well under a tenth of one percent of the variance in test scores. To call such an effect "substantial" strains credulity.

Kevin Welner
Project Director

William Mathis
Managing Director

Jennifer Berkshire
Academic Editor

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

GREAT LAKES CENTER
FOR EDUCATION RESEARCH & PRACTICE

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 *Urban Charter School Study 2015*

Andrew Maul, University of California-Santa Barbara

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.

The overall thrust of these reports (as well as the literature on charter schools in general) has been that there is essentially zero difference in overall performance between demographically-similar students in charter schools and in TPSs. Given this, a reasonable next step is to explore whether larger differences can be found for particular subgroups of schools, and what might explain any such differences. CREDO's latest study, the *Urban Charter School Study*,[1] pursues such a tactic by investigating charter schools in urban environments. Using a methodological approach highly similar to CREDO's previous reports, the study concludes that, overall, charter schools in urban environments provide a slightly greater test score advantage to students than those in non-urban environments. Specifically, students in urban charter schools were estimated to score approximately 0.055 standard deviations higher on math tests and 0.039 standard deviations higher on reading tests than their peers in urban TPSs.

II. Findings and Conclusions of the Report

The main conclusions presented in the report are as follows:

- On average, it was estimated that students in charter schools in the 41 urban areas fare better on academic tests relative to their peers in "feeder" traditional public schools (i.e., the TPSs that charter students attended prior to transferring into charter schools) in the same areas. Charter students were estimated to score 0.039 standard deviations higher on reading tests and 0.055 standard deviations higher on math tests relative to their TPS peers. Given that these findings appear to be

derived from the same dataset used in the 2013 national study, which found near-zero overall differences (i.e., differences of less than 0.01 standard deviations) in both reading and math, this would appear to imply that students in TPSs outperform students in charter schools in non-urban environments, though this point is not discussed in the report.

- The apparent advantage of charter school enrollment was estimated to be slightly greater on average for Black, Hispanic, low-income, and special education students.
- As with the 2013 report, the apparent advantage for charter schools appeared to increase slightly overall during the time period considered (between the 2008-09 school year and the 2011-12 school year).
- 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.
- There were significant region-to-region variations in the estimated differences between charter and TPSs.

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

The conclusions of the study are based primarily on analyses of datasets furnished by the 22 state departments of education, which are collectively stated to include observations from 1,018,510 charter school students along with a matched group of comparison students from "feeder" traditional public schools, covering the 2006-07 through 2011-12 school years.

Overall, the study concludes that "urban charter schools on average achieve substantially greater levels of growth in math and reading relative to local TPS" (p.43). There are no explicit policy recommendations stated in the report, though a variety of "implications" are explored, such as the following:

> . . . the results . . . provide ample evidence that some urban charter sectors have figured out how to create dramatically higher levels of academic growth to their most disadvantaged students. . . . These urban regions can serve as models from which all public schools serving disadvantaged student populations may learn (p.38).

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

In earlier reviews, Miron and Applegate,[2] Maul[3], and Maul and McClelland[4] have called attention to a variety of technical and conceptual concerns with the methods employed by the CREDO charter school studies. For the most part, it does not appear that CREDO researchers have altered their methodology in light of those concerns; thus, many of the comments made here overlap with previously raised issues. Of particular concern is the approach used to match students to "virtual twins" for comparison, and the reporting of year-to-year changes in test scores in terms of "days of learning." Additionally, a number of choices made in the analysis of data and reporting of results are insufficiently described and justified.

Concerns about the Matching Procedure

Defending a causal inference in the absence of a controlled experimental design, in this instance, means that observational data can be used to provide an estimate of what would have happened to charter school students had they attended a traditional public school. CREDO's approach to this estimate is 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.

VCRs are a home-grown technique of CREDO's. The choice to commit to such a technique is concerning given the availability of and more broadly used propensity-based score matching (PSM) techniques,[5] which appear to be superior in several respects. First, the VCR technique requires exact matches, whereas propensity-based methods do not, meaning that arbitrary cutoffs for continuous covariates (e.g., matching to twins with a prior year test score within 0.1 standard deviations, as is done in the VCR technique) are unnecessary. Second, the VCR technique found a match for only "greater than 80%" of charter students (Technical Appendix, p.8), meaning that close to 20% of charter students were excluded from the study. In the 2013 report it was indicated that the excluded students had average scores 0.43 standard deviations lower than the average of the included students, introducing the potential for bias due to systematic exclusion of lower-performing students. (A propensity-based method would probably have allowed inclusion of far closer to 100% of the charter sample.) Third, and more broadly, by committing to an

in-house technique like VCRs instead of using more established methodology—and, further, by refusing to explain the rationale for this choice—CREDO has made it difficult for other researchers to evaluate the adequacy and appropriateness of the study's methods. For example, evaluating the success of a matching procedure by checking for baseline equivalence (*i.e.*, comparability) in matched groups (as a safeguard against selection bias) is *de rigueur* in the PSM literature, but this step is apparently skipped in the present study.

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 (*i.e.*, gender, ethnicity, English proficiency status, eligibility for subsidized meals, special education status, grade level, and prior test scores),[6] the only remaining meaningful difference between students is their school type. This requires, essentially, a leap of faith. One must believe, for example, that the dichotomous "eligibility for subsidized meals" variable (along with the other aforementioned demographic variables) is sufficient to control for all meaningful socioeconomic differences in students. Additionally, it seems plausible that "selection effects" could be at play. That is, there may be systematic differences in the overall extent to which parents of charter school students are engaged with their children's education (given that such parents are by definition sufficiently engaged to actively choose to move their children to a charter school), and one must simply believe that the seven aforementioned demographic variables control for all such differences.

Setting aside all concerns over methods, the actual effects reported in this study are fairly small in magnitude, and should not be given more weight in policy considerations than they deserve.

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.[7]

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. There are significant controversies in the psychometric literature regarding the relationship between learning and test scores; without a clear (or indeed any) rationale for this choice of metric, the expression of findings in terms of "days of learning" cannot be regarded as credible.

Furthermore, as Miron and Applegate[8] 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 multilevel modeling; it is unclear and puzzling why this was not done.

Arbitrary and Unexplained Analytic Choices

A number of choices made in the analysis and reporting are given an insufficient amount of explanation to permit the reader to judge their appropriateness. Of particular concern are the following:

- The details of the procedure for selecting urban regions and schools within those regions are never given, beyond a statement that the "factors considered include[d] total city population, total population of charter students, total number of urban charter students, size of the primary school district(s), and charter market share," and that this was "cross referenc[ed] . . . with available data" (Technical Appendix, p.14).

- The details of selecting "feeder schools" are not given. In fact, it is never described in either the main report or the Technical Appendix that the "virtual twins" are not drawn from the general population of traditional public schools, but rather, only from the subset of such schools from whom students leave to enroll in charter schools. Thus the results of the study cannot be interpreted as head-to-head comparisons of charter schools and TPSs, but rather, at best, comparisons of charter schools and the specific subset of TPSs from which they draw students (which, plausibly, may have average performance levels significantly below the mean of TPSs in general). Although this is arguably a reasonable choice, its lack of discussion invites the misinterpretation that charter schools are being compared to TPSs in general.

- When multiple tests of statistical significance are conducted using the same dataset, there is an increased chance of committing a Type I error (or "false positive"), i.e., mistakenly declaring an effect to be real when it is in fact due only to random chance. No correction is applied to control for this. This issue is given some attention on pp.17-18 of the Technical Appendix, though the rationale is difficult to understand. For example, it is stated that a Bonferroni correction "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)" (p.17). However, this does not explain why an inflated Type I error rate is an ignorable concern (and there are many corrections available that do not make the same strict assumptions as the Bonferroni correction). Later in the passage, it is stated that "small effect sizes . . . 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." This is simply a statement of preference: unsurprisingly, the study's authors would rather err on the side of false positives than false negatives (or, in other words, they would

rather err on the side of over-claiming rather than under-claiming). The lack of a correction leaves open the possibility that many of the study's results are, in fact, due to chance error.

- No correction is applied for the fact that the data are hierarchical (in the sense that students are nested within classrooms and schools), violating a key assumption of parametric statistics (i.e., independence of observations). There may 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. It is stated that

  > . . . the decision was made not to cluster errors at the school level . . . due to the existence of urban regions that, while containing a substantial number of students in total, nonetheless had a large number of schools with relatively small student bodies . . . clustering standard errors at the school level reduces aggregate statistical power to a degree that more than offsets the benefit of estimating standard errors at the school level (Technical Appendix, pp.10-11).

  This claim is simply untrue; there is a robust literature on the correction of standard errors for clustering, regardless of the within-cluster sample sizes. This further adds to the possibility of false positives due to chance error.

- For the most part, "charter schools" are presented as a monolithic category; little consideration is given to factors such as whether a charter was granted by a school district or another entity, operated by a large corporation, had large differences in personnel policies, or the like.

- As before, no form of test-equating is employed, thus assuming that all scores can be standardized and projected onto a common metric.[9]

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 *Urban Charter School Study*. However, even setting aside all of these concerns, the actual effects reported in this study are fairly small in magnitude, and should not be given more weight in policy considerations than they deserve. The overall effect sizes reported are 0.039 standard deviations for reading tests and 0.055 standard deviations for math tests. If they were correct, these numbers could be interpreted as stating that well less than a tenth of one percent of the variation in test scores can be attributed to whether a student is in a charter school or a "feeder" traditional public school. Calling such an effect "substantial" (p.43) strains credulity. 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. Thus, while the effect sizes reported for urban charter schools are marginally higher than the near-zero effect sizes reported for charter schools as a whole in the 2013 study, the magnitude of this difference may be interpreted as trivial.

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 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; additionally, it is valuable to explore possible trends in the effectiveness of schools related to factors such as whether the school is located in an urban environment. However, this review has noted a number of concerns with the methodology and reporting of CREDO's study. As such, the findings of this report cannot be regarded as compelling evidence of the greater effectiveness of charter schools compared with traditional public schools, either overall or specifically within urban districts.

Notes and References

1 Center for Research on Education Outcomes (CREDO) (2015, March). *Urban Charter School Study*. Palo Alto: CREDO, Stanford University. Retrieved April 22, 2015, from http://urbancharters.stanford.edu/index.php.

2 Miron, G. & Applegate, B. (2009). *Review of "Multiple choice: Charter school performance in 16 states."* Boulder, CO: National Education Policy Center. Retrieved April 22, 2015, 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 April 22, 2015, from http://nepc.colorado.edu/thinktank/review-charter-performance-michigan.

4 Maul, A., & McClelland, A. (2013). *Review of "National Charter School Study 2013."* Boulder, CO: National Education Policy Center. Retrieved April 22, 2015, from http://nepc.colorado.edu/thinktank/review-credo-2013/.

5 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.

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 A standard-deviation-to-days-of-learning crosswalk table is given on p.11 of the report, revealing that (for example) what is usually considered a "small" effect size of 0.20 translates to 144 "days of learning," but the conversion procedure and its rationale are never discussed.

8 Miron, G. & Applegate, B. (2009). *Review of "Multiple choice: Charter school performance in 16 states."* Boulder, CO: National Education Policy Center. Retrieved April 22, 2015, from http://nepc.colorado.edu/thinktank/review-multiple-choice.

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

Miron, G. & Applegate, B. (2009). *Review of "Multiple choice: Charter school performance in 16 states."* Boulder, CO: National Education Policy Center, 6. Retrieved April 22, 2015, from http://nepc.colorado.edu/thinktank/review-multiple-choice.

Document Reviewed:	**Urban Charter School Study Report on 41 Regions 2015**
Author:	Center for Research on Education Outcomes (CREDO)
Publisher/Think Tank:	CREDO
Document Release Date:	March 2015
Review Date:	April 27, 2015
Reviewer:	Andrew Maul, University of California, Santa Barbara
E-Mail Address:	amaul@education.ucsb.edu
Phone Number:	(805) 893-7770

Suggested Citation:

Maul, A. (2015). *Review of "Urban Charter School Study 2015."* Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/review-urban-charter-school.