EPIC Education and the Public Interest Center
University of Colorado at Boulder

EPRU Education Policy Research Unit
Arizona State University

DOCUMENT REVIEWED:	"Multiple Choice: Charter School Performance in 16 States"
AUTHOR:	Center for Research on Education Outcomes
PUBLISHER:	Center for Research on Education Outcomes (CREDO), Stanford University
DOCUMENT RELEASE DATE:	June 15, 2009
REVIEW DATE:	June 24, 2009
REVIEWERS:	Gary Miron and Brooks Applegate
E-MAIL ADDRESS:	gary.miron@wmich.edu, brooks.applegate@wmich.edu
PHONE NUMBER:	(269) 599-7965

SUGGESTED CITATION:
Miron, G. & Applegate, B. (2009). *Review of "Multiple choice: Charter school performance in 16 states."* Boulder and Tempe: Education and the Public Interest Center & Education Policy Research Unit. Retrieved [date] from http://epicpolicy.org/thinktank/review-multiple-choice

Summary of Review

The Center for Research on Education Outcomes (CREDO) at Stanford University conducted a large-scale analysis of the impact of charter schools on student performance. The center's data covered 65-70% of the nation's charter schools. Although results varied by state, 17% of the charter school students have significantly higher math results than their matched twins in comparable traditional public schools (TPS), while 37% had significantly worse results. The CREDO study strengthens the well-established, broader body of evidence showing average charter performance to be equal to, or perhaps lower than, the performance of traditional schools—a body of evidence that is summarized in this review. The study also presents some state-level analyses concerning policy options; this review points out limitations with those analyses and also explores other policy implications of the report's findings. The relative strength and comprehensiveness of the data set used for this study, as well as the solid analytic approaches of the researchers, makes this report a useful contribution to the charter school research base. Nevertheless, this review points out some weaknesses and areas for improvement, many of which represent commonplace limitations for this type of study that should be shared in the technical report.

Review

I. INTRODUCTION

In recent years, several attempts have been made to draw overall conclusions regarding charter school performance from multiple states or from multiple studies.[1] These efforts seek to inform broader policy decisions regarding whether charter school policies are likely to generate schools that are more effective than traditional public schools. The new study from the Center for Research on Education Outcomes (CREDO) at Stanford University, *Multiple Choice: Charter School Performance in 16 States*,[2] goes beyond earlier efforts in its attempt to answer the broader policy question about the relative performance of charter schools.

Charter schools, by design, receive more autonomy in operations; in exchange they are to be held more accountable than other public schools for student outcomes. Charter leaders use this autonomy to create their own schools, select their own governing boards, design educational interventions appropriate for students' unique needs and learning styles, and hire and fire teachers more freely. In turn, the enhanced autonomy granted to charter schools was expected to result in, among other things, greater performance of students enrolled in them.[3]

Performance accountability was to be effected through market mechanisms (with funding following students) and contractual relationships with authorizers. If parents were not satisfied they would not enroll or they would choose to leave, which could eventually bankrupt the school. If the authorizer (e.g., a state board of education or a local school district) did not believe the school was living up to its mission and agreed-upon contract, the contract or charter could be revoked or not renewed after it expired. When poorly performing charter schools are removed from the ranks, the aggregate results of the remaining charter schools would go up. (In contrast, one would expect average results from traditional public schools, which are not easily closed, to be weighted down by a number of poorly performing schools.) These sorts of accountability mechanisms also suggest why charter schools were, on average, expected to outperform traditional public schools.

Research on charter performance, such as the new CREDO study, is intended to evaluate whether or not charter policies are fulfilling policymakers' intentions. The scope and relative rigor of the CREDO study reinforces the larger body of evidence, which shows no overall impact of charter schools on performance.

II. FINDINGS AND CONCLUSIONS OF THE REPORT

Findings of the CREDO report are stated to be based on a rigorous analysis of more than 70% of the nation's students attending charter schools. The key findings include the following:

- Charter school students on average saw a decrease in their academic growth in reading by 0.1 standard deviations and 0.3 standard deviation units for math. These decreases are small but statistically significant.
- For example, 17% of the charter school students have significantly higher math results than their matched twins in comparable traditional public schools (TPS),

while 37% had significantly worse results. Forty-six percent of the charter school students had math gains that were statistically indistinguishable from the average growth among the comparison TPS students.

- 5 states had significantly higher learning gains for charter school students than would have occurred in traditional schools: Arkansas, Colorado (Denver), Illinois (Chicago), Louisiana, and Missouri.
- 6 states had significantly lower average charter school student growth than their TPS matched peers: Arizona, Florida, Minnesota, New Mexico, Ohio, and Texas.
- 4 states had results that did not differ from the traditional public schools: California, District of Columbia, Georgia, and North Carolina.
- States with caps on the number of charter schools tended to have lower academic growth, as did states that had multiple authorizers. Evidence also suggested that states with an appeal process to overturn denied charter applications had a small but significant advantage in terms of growth in student achievement.
- Charter students in elementary and middle school grades had small but significantly higher rates of learning than their matched peers in traditional public schools, but students in charter high schools and charter multi-level schools had significantly worse results.
- Gains in achievement scores were lower for African American and Hispanic students enrolled in charter schools than for their matched peers in traditional public schools.
- Charter schools were found to have slightly better academic growth results for students in poverty and for students who are classified as English Language Learners.[4]
- Students in special education programs had similar outcomes as their matched peers in traditional public schools.
- Charters in their first year showed results that ranged from poor to very poor.[5]

III. THE REPORT'S RATIONALE FOR ITS FINDINGS AND CONCLUSIONS

This is an empirical report. The findings and conclusions are based on a longitudinal student-level data set created by the researchers at CREDO. This dataset reportedly includes student-level data from 16 states (including D.C.), although Illinois and Colorado data are comprised of charter school students in only Chicago and Denver, respectively.

Although full details on the methods are not included, it is apparent that conclusions are based solely on the findings from the analysis of the student-level data set.

Conclusions about policy implications are based on a more rudimentary classification of states, with regard to (i) whether they have caps on the numbers of charter schools, (ii) whether they have multiple authorizers, and (iii) whether they have an appeal process. A discussion of these policy implications is included later in this review.

IV THE REPORT'S USE OF RESEARCH LITERATURE

The contents of the report focus on its new findings. Minimal attention is given to the existing body of literature on student achievement in charter schools, and little effort was made to link the new findings to the body of evidence.

The research base on charter schools has clearly improved over time. In the mid-1990s, this research largely focused on start-up issues and the degree to which these

schools were innovative or the extent to which they promoted racial integration or segregation. By the end of the 1990s, more of the evaluations of charter schools addressed student achievement.

In 2001, Miron and Nelson[6] prepared a synthesis of evidence on student achievement and charter schools and found 15 studies of charter school achievement across 8 states. Half of these were of limited scope, meaning they included few schools and years of data. Using similar selection criteria for studies in 2008, Miron, Evergreen, and Urschel[7] identified and synthesized the evidence from 47 studies. While this represented a large increase in the number of studies, many of these were still of limited scope.[8]

National Studies of Student Achievement in Charter Schools

The press release for the CREDO study makes the claim that this is the first national study of charter school impact. Although the claim is overstated, there is an element of truth. Prior to the CREDO study, there were a number of studies that truly encompassed charter results across the nation (see, for example, Nelson, Rosenberg and Van Meter, 2004; Hoxby, 2004; Lubienski & Lubienski, 2006; Braun, Jenkins & Grigg, 2006).[9] However, all of these national studies were based on cross-sectional designs, which meant they could describe the relative status or performance of charter schools, but they could not measure impact or change over time. Except for the Hoxby report, which used school level data from state assessment programs, these studies used student-level data from the National Assessment of Educational Progress (NAEP). Also except for Hoxby, all of these national studies found that charter school students had test results similar to or worse than comparable traditional public school students. The results from the CREDO study are similar to the earlier national studies, although the CREDO study goes beyond these earlier ones by using student-level data, with a large (almost national) scope, to look at changes in charter schools over time.[10]

Multi-State Studies of Charter Schools

The CREDO study is a large, multi-state study that claims to cover 15 states plus the District of Columbia (DC). In fact, the results indicate that Massachusetts was left out, and thus only 14 states and DC are included.[11]

There have been earlier efforts to combine findings from multi-state studies. Using school-level data, Loveless (2003)[12] combined achievement results in 10 states and compared these with traditional public schools. Miron, Coryn, & Mackety (2007)[13] similarly used school-level data to conduct a six-state study of the Great Lakes states, which covered one-quarter of the nation's charter schools. The Loveless study used Z scores, similar to the CREDO study, to combine and compare across states. The Great Lakes study used residual scores. The Loveless study found that charter schools were performing at lower levels than were traditional public schools, but they were improving more rapidly than traditional public schools over time. The Great Lakes study had similar results. However, that study also found that as performance levels of charter schools in more mature states improved and approached the level of demographically similar traditional public schools, they tended to level off and remain similar to the performance level of traditional public schools. Only in Illinois did the charter schools performance level eventually surpass their comparison group.

The results from the CREDO study are simi-

lar to those from these earlier multi-state studies, although the CREDO study goes beyond these earlier studies because it is based on student-level data rather than school-level data. This is an important quality distinction.

Longitudinal Studies with Student-Level Data

The researchers at CREDO have made an exceptional effort to secure student-level data from so many states. Most past studies of student achievement in charter schools employ longitudinal designs with *school*- or *group*-level data.[14] Although these studies are the most prevalent, they are not as reliable as studies that can examine impact based on student-level data. Since 2001, however, an increasing number of charter school studies are based on individual student data, which can link test results for students over time to measure gains or changes in performance. Student-level data allow analysts to match charter and non-charter students on a set of demographic characteristics and then track their relative performance over time. The first study of student achievement in charter schools that used a matched student design focused on Arizona and was completed in 2001 (see Solmon, Paark, & Garcia, 2001).[15] Since then several studies have been based on student-level data, including ones for California (Zimmer, et al., 2003), Delaware (Miron, et al., 2007), Idaho (Ballou, Teasley, & Zeidner, 2008). Florida (Sass, 2006), North Carolina (Bilfulco & Ladd, 2006), and Texas (Gronberg & Jansen, 2005).[16]

A more recent study by RAND (2009) used student-level data and covered eight states.[17] Matched student designs are the most promising development in research on student achievement in charter schools.[18] The costs are reasonable low and one can conduct large-scale studies with relatively strong controls. The new CREDO study is a positive illustration of what can be done when states are willing to grant researchers access to student-level data sets.

Randomly Controlled Experiments

Although the CREDO study is both rigorous and comprehensive, there are studies that have employed designs with, at least in terms of internal validity, even greater rigor. These include a number of smaller-scale studies that have simulated random assignment by creating control groups from admissions wait lists (see Hoxby and Rockoff, 2004, Hoxby, Muraka, 2007, and Abdulkadiroglu et al., 2008).[19] These studies have shown positive effects for the particular charter schools studied, although they lack external validity since they include a small number of the higher-performing charter schools, with sufficiently large waiting lists, that are willing to participate in such studies.[20] That is, the results cannot validly be generalized to less-popular charter schools.

A true randomly controlled experiment will be difficult and expensive and will come with its own set of obstacles. However, the U.S. Department of Education has indeed funded a large and expensive randomly controlled study of student achievement in charter schools, which is being conducted by Mathematica Policy Research Inc. This study will only cover 40 charter schools, but comes with a price tag of more than $5 million.[21] Although randomly controlled experiments are commonly perceived as the gold standard for study design, they are very difficult to implement when evaluating policies as broad as charter school reforms. The fact that not a single such study has yet been successfully completed on charter schools underlines the difficulty involved in implementing such designs when evaluating large-scale reforms.

Until we know whether the Mathematica study can be successfully completed, the CREDO study stands as the relatively most rigorous and comprehensive study of student achievement in charter schools.

Meta-analyses and Syntheses of Evidence

A number of efforts have been made to synthesize evidence across studies. In 2001, Rand researchers[22] summarized the evidence across 3 studies, and Miron and Nelson summarize the evidence across 12 studies of student achievement in charter schools. More recent summaries have been prepared by Carnoy et al. (2005), Hill (2006), and the National Alliance for Charter Schools (2009).[23] The later three examples grouped studies depending on whether they were positive or negative, and in some cases they are grouped by design type, although no effort has been made to weigh and synthesize the results across studies.

The most rigorous attempt to synthesize the evidence on charter school impact was undertaken by Betts and Tang (2008). They used meta-analysis methods to combine and synthesized findings across 14 studies that covered 7 states and 2 school districts. They only included more rigorous studies that used student-level data. The median effect size across these studies was barely distinguishable from zero (0.005). A recent effort by researchers at Western Michigan University[24] synthesized the evidence across 47 studies. The findings across the states were distinguished by quality, scope, and the nature of their impact. Overall, 19 studies had positive findings, 12 studies had mixed findings, and 16 had negative findings. The mean impact rating for charters was indistinguishable from zero.

We have included this short review of the literature to supplement the CREDO report and to illustrate that although results vary within and across states, the overall answer for policymakers regarding the impact of charter schools was well established prior to the new report and was indeed reinforced by the new findings: on the whole charter school students are performing similar or slightly worse than comparable students in traditional public schools. This finding has not changed over time, nor has it changed as the body of evidence has expanded to include more states and more rigorous studies. The findings from the CREDO study simply strengthen this overall conclusion.

V. Review of the Report's Methods

The CREDO study undertook an incredibly ambitious goal of securing data-sharing agreements with 16 charter-school states to obtain student-level data that could be longitudinally linked. However, the researchers' next task, developing a common data structure necessary for merging data and performing analyses, was potentially more problematic than the first. The complexity of this task warrants specific details as to how the common data structure was constructed. Unfortunately, these details were not provided, even in the technical appendix. In addition, the report shows a considerable lack of clarity regarding the intended unit of analysis. Given the title, and given much of the discussion of findings, we expected the unit of analysis to be charter schools, but all of the analyses presented in the report are based on student-level data.

Four Technical Concerns

Notwithstanding our judgment that this report offers a great deal of useful information and analyses, we observed four issues with the analysis that should be noted. All four concern methodological choices and report-

ing of information, so we ask for patience from readers not well-versed in assessment and statistics.

Construction of a Meaningful Outcome Variable

In the process of creating a dependent variable that represents a valid presentation of student achievement across 16 (or 15) states CREDO researchers appear to have confused score standardization and test equating. In constructing the dependent variable, these researchers have, at a minimum, assumed that test equating is not needed to form a valid dependent variable. In fact, the CREDO researchers simply projected the student performance from each different state achievement test onto a common scale and then interpreted all the scores on this scale, as if the scores all mean the same thing. This represents a common error in psychometric score interpretation.[25]

The Virtual Twin as a Match

The "virtual twin"[26] matching method CREDO created is similar to a matching technique employed by the two authors of this review in their state evaluation of charter schools in Delaware.[27] Similar to that evaluation, exact matches were established for students in a charter school. Unfortunately, it is not clear in either the report or its technical appendix why the previous test score (t-1) was matched by ±0.1 range and not by a propensity method, an increasingly common matching method for continuous response variables. Furthermore, exact matching efficiency is a function of the size of the feeder school population, and the CREDO report did report the overall percentage match at the level of the state. Unfortunately these researchers did not provide the accuracy of the matching procedure at the level which it was preformed (e.g., the feeder school level).[28]

Regarding the suitability of the feeder school logic, CREDO researchers could have assembled all of the virtual twins into a "virtual school" and examined the comparability of the virtual schools to all of the surrounding feeder schools. In a weak way this can give an indication of the extent of selection bias inherent in the charter school population that has been captured by the matching procedure. If the virtual school looks similar to the feeder schools, then (as measured by the matching variables) there would not be a substantial amount of selection bias operating.[29] Lastly, it is important to remember that matching students does not create matched schools. There is more to a school than its students, and comparing students is not equivalent to comparing schools.

Primary Analysis: OLS Regression with Robust Standard Errors

To address the known and probably nontrivial intra-class correlation among students nested in the same school and schools nested in states, CREDO researchers used robust standard errors on ordinary least squares (OLS) estimates. While this technique represents a vast improvement over conventional OLS standard errors, there is considerable detail omitted that a technically based reader would want in order to understand how this was implemented.[30]

Year-to-Year Gain Scores

The primary findings presented in the CREDO report are based on an average year-to-year gain score expressed in standard deviation units. Although the technique used by CREDO researchers is not unusual, it does not reflect a true longitudinal growth of individual student achievement. Rather it

represents average growth of a group.

This again reflects possible confusion of the unit of analysis. If the group remains relatively stable, then weak inferences can be derived about the group's relative improvement or decline in achievement. However, if the stability of the group changes, then even weak inferences about group gains or losses cannot be made. This is because changes in the group become confounded with simultaneous changes in time. Unfortunately, there was likely instability in the "virtual twin" group that can only weaken the inferences drawn from these analyses. Moreover, the validity of such an approximation of longitudinal gain is stretched to the limits if even a small sub-sample of schools uses year-around-schooling. Thus, it is not clear—and accordingly, justification is needed to understand—why consecutive year-to-year gain scores were averaged.

VI. REVIEW OF THE VALIDITY OF THE FINDINGS AND CONCLUSIONS

Overall the findings reported by CREDO parallel the increasing body of research related to the impact of charter schools on student achievement. As previously presented, there is much variation among these studies. This report adds to that variation by creating a multi-state, student-level matched design. Additionally, between three and eight years of data were pooled, depending on the state. Thus the potential for the CREDO report to inform researchers and policymakers on the effects of charter school attendance on student achievement was promising. Unfortunately, there are a number of methodological limitations that weaken the CREDO conclusions. One notable example is that the report includes little explanation for the use of averaged gain scores when true longitudinal growth could have been examined.[31] More generally, it would have been very helpful for the report or its technical appendix to have included more details and justification for decisions regarding study design and methods of analysis.

Further, we see four notable threats to the validity of the report's core findings. First, the CREDO analyses are grounded in the assumption that the clustering effects of state and school were minimal and were effectively attenuated by the use of robust standard errors. Yet there is no assurance that the use of robust standard errors adequately accommodates the nested structure of the data. Unfortunately, the method used for matching precluded use of hierarchical linear modeling (HLM), which may have been more appropriate for these data.

Second, the analyses are also grounded in the assumption that averaged groups' gain scores represent the longitudinal growth trajectories of individual students. This is a weak proxy for individual student longitudinal growth curves.[32] Students should be nested in schools and, in a multi-state data set, schools should be nested in states. For example, there are notable differences in state chartering laws, such as caps that have nontrivial impact on the ability of schools to affect student achievement, as argued in this report. In the present CREDO analysis, this natural nesting of data is ignored. As a consequence, the degrees of freedom in the model hypotheses tests are extremely large, which makes even trivial differences "statistically significant."

This brings us to the third point regarding the validity of the findings. As can be seen from the numerous tables presenting the results of the OLS regression models, the sample sizes were extremely large (>1.7 million). Although there were often more than 40 predictor variables, the sheer size of the sample ensured the mean square error

would have more than 1.5 million degrees of freedom. Thus, every predictor will be statistically significant. In other words, declaring an effect as statistically significant when they all are, or will be; regardless of the effect size, may be overstating the findings. Many of the graphs present extremely small differences in achievement (e.g., .01 and .03 SD units; p. 22) between charter schools and traditional public schools, representing a .0039 and .0119 percentile difference if one assumes a normal distribution. The CREDO researchers should revisit the *practical* significance of their findings, tempered by the numerous assumptions and limitations inherent in their methods.

Our last concern focuses on the multivariate nature of the two dependent variables (reading results and math results). It cannot be assumed that performance on reading and mathematics assessments are not correlated,so any analysis of reading is partially an analysis of math (and vice versa), unless each is partialled from the other.[33] This could easily have been done in the CREDO study by simply adding the non-dependent variable into the model as a predictor. For example, if math were the dependent variable, then the first predictor would be reading score. This would effectively partial out the covariance between math and reading, so the effects of the remaining predictor variables could be interpreted for their ability to explain variation among averaged math gain scores.

We offer one final note of concern to readers attempting to ascertain the validity of the report's key conclusions. The report's technical appendix includes results from several OLS models, presented in tables. In many cases the predictor variables can be understood (e.g., "is English Learner, " p. 4), but for other variables it is not clear how they were coded. For example, the title of the report indicates an analysis of 16 states, yet only 15 states are represented in the OLS models (Massachusetts is missing). Thus, one is tempted to interpret this as the referent state where the remaining 15 states are dummy coded against it. If so, what is the justification for choosing Massachusetts as the referent? This explanation should be included in the report or its technical appendix. Moreover, Grade05 is omitted in the OLS models, suggesting that this was the referent group for grade. But again, no explanation is provided.

Effect of Charter School Policy on Performance

In a section of the CREDO report titled "Effects of State Charter School Policy on Performance," the researchers explore the relationship between their state-level findings and three policy issues: (i) presence of caps on the number of charter schools, (ii) existence of multiple authorizers, and (iii) existence of an appeal process. The researchers concluded that states had significantly lower achievement growth if: (a) they had caps on the number of charter schools that could operate, or (b) they had multiple authorizers. Evidence presented in the report also suggested that states with an appeal process to overturn denied charter applications had a small but significant advantage in terms of growth in student achievement.

However, given the manner in which states define caps, and given the wide range of practices when it comes to enforcing those caps, any analysis of this variable cannot be expected to reflect valid findings.[34] The same is true of appeals, given the wide range of practices when it comes to using the appeal process. Even the issue of multiple authorizers is clouded by distinct variations in the activity of those authorizers. Rather than look at how a state's charter school legislation is worded, it would be more useful to

base an analysis on how the legislation is being interpreted and implemented. Caps, multiple authorizers and the existence of an appeals process are all viewed as components of less restrictive charter school laws that would spur the opening of larger numbers of charter schools. (And note that the purported effects of the three factors here are inconsistent in this regard.) A better way to view whether a charter school reform is restrictive or permissive is to look at the number of charter schools in operation as well as the relative speed with which the reform has grown over time.

We created Table 1 to sort the CREDO states into three groups, depending on whether their charter schools students had (a) significant positive growth relative to their virtual twins in traditional public school, (b) significant negative growth relative to their virtual twins, or (c) no difference in growth relative to their virtual twins. The states in the positive group have relatively fewer charter schools, and many of the largest charter schools states had negative findings.

The group of more successful charter school states has only 61.6 schools on average, while the unsuccessful states have 275 charter schools on average. The ranking next to each state name indicates the relative rank of the state in terms of the number of charter schools operating in 2008. There are 41 charter school laws; thus, the ranking runs from the largest charter school state, California, to the smallest charter school state, Mississippi, ranked #41 with only one charter school.

The relationship between the size of a state's charter school reform and its relative performance, as measured by the CREDO researchers, is striking. It suggests that states with fewer charter schools are better able to

Table 1.
Number of Charter Schools per State

		Total CS in 2008	*Group Mean*	*Group Medium*
States with positive Charter school findings				
Arkansas	#33	18		
Colorado	#9	140		
Illinois	#61	60		
Louisiana	#21	54		
Missouri	#23	36		
subgroup total		**195**	**61.6**	**54.0**
States with negative charter school findings				
Arizona	#2	479		
Florida	#3	348		
Minnesota	#8	148		
New Mexico	#15	66		
Ohio	#5	295		
Texas	#4	314		
subgroup total		**1650**	**275.0**	**304.5**
States with no significant differences				
California	#1	703		
District of Columbia	#14	74		
Georgia	#16	65		
North Carolina	#11	103		
subgroup total		**945**	**236.3**	**88.5**

Note: The actual number of charter schools for Colorado and Illinois considered in the study include only those schools in Denver and Chicago, respectively.

Source: National Center for Education Statistics, Table 4.3 Charter school legislation by state: 2008.

oversee those schools and hold them accountable. In fact, the existence of caps in many states has been an instrument to exert pressure on authorizers to close poor-performing charter schools in order to create places for new applicant groups.[35]

Related to overall size of the state charter school reforms is the rate or speed of growth. A sensible way to implement any new school reform is to begin implementa-

tion on a small scale, to ensure that things are working as anticipated.[36] It is not surprising, then, that the states in the low-performing group, with the possible exception of Minnesota, ramped up their charter school numbers very quickly. Therefore, even states like Ohio that had good measures of accountability written into its law found that it was not practically possible to track and oversee the plethora of schools that were being launched.

Another factor that the CREDO researchers might consider is the relationship between state outcomes and the extent to which the charter schools are operated by private education management companies (EMOs). Table 2 illustrates this relationship. As is apparent, the poor-performing states had 21% of their schools run by for-profit EMOs, while the high-performing states only had 13.7% run by these for-profits. Interestingly, the successful states did have slightly more schools run by nonprofit EMOs (20.5% compared to 18.2% for the low performing states). The proportion of nonprofit EMOs among the high-performing states was particularly bolstered by Illinois (71.6% of all charters), where nonprofit EMOs are being used as a means of scaling up or expanding the number of high-performing charter schools.

Aside from looking at the actual size of the state charter school reforms and the prevalence of EMOs, other important variables that could be considered in further analyses by the CREDO researchers include the rigor of the application process and the rigor of oversight. A framework for ranking charter school laws developed by Chi and Welner (2008)[37] and an AERA paper by Miron (2005)[38] both outline a number of variables related to strong charters schools, as opposed to permissive charter school laws.

Table 2.
Proportion of Charter Schools Operated by For-Profit or Nonprofit EMOs

	For Profit EMOs	*Nonprofit EMOs*
States with positive charter school findings		
Arkansas	11.1%	27.8%
Colorado	10.7%	1.4%
Illinois	12.2%	71.6%
Louisiana	7.4%	9.3%
Missouri	38.9%	2.8%
subgroup total	**13.7%**	**20.5%**
States with negative charter school findings		
Arizona	22.3%	21.5%
Florida	36.2%	2.9%
Minnesota	4.1%	0.7%
New Mexico	0.0%	0.0%
Ohio	31.5%	22.4%
Texas	4.8%	38.5%
subgroup total	**21.0%**	**18.2%**
States with no significant differences		
California	2.4%	15.9%
District of Columbia	10.8%	27.0%
Georgia	12.3%	3.1%
North Carolina	4.9%	1.9%
subgroup total	**4.0%**	**14.4%**

Source: The data for this table are derived from the national profile reports of For Profit and Nonprofit EMOs: http://epicpolicy.org/by-topic/publications/732.

VII. USEFULNESS OF THE REPORT FOR GUIDANCE OF POLICY AND PRACTICE

The CREDO researchers noted that this report would be followed by two additional reports. The relative strength and comprehensiveness of their data set, as well as the relatively solid analytic approaches of the researchers, makes this first report a useful contribution to the charter school research

base. The two future reports should add further information that will be useful for policymakers. This review does point to some weaknesses and areas for improvement, many of which represent limitations—not outside the range of limitations that are inherent in other studies on student achievement in charter schools—that should be shared in the technical report. The review offers suggestions that we hope will help improve the subsequent reports produced by CREDO based on this multi-state data set.

Notes and References

[1] These studies are described and cited in Section IV of this review.

[2] Center for Research on Education Outcomes (CREDO) (2009, June). *Multiple Choice: Charter School Performance in 16 States*. Palo Alto: CREDO, Stanford University.

[3] Chi, W. C., & Welner, K. G. (2008). Charter Ranking Roulette: An Analysis of Reports that Grade States' Charter School Laws. *American Journal of Education*, *114*(2), pp. 273-298.

[4] This is a surprising finding, given that African American and Hispanic students in charter schools had lower results than their matched twins in traditional public schools. This finding could be due to suppressor effects operating in the OLS regression model. Thus, CREDO researchers should seek to determine if these findings are as they appear or are due to suppression effects.

[5] Unfortunately, it is not completely clear whether this variable was included in the OLS models. This may refer to the variable: "fall of starting score test year." More detail in the technical appendix would help clarify such questions. In any case, the CREDO researchers cite possible explanations for this first-year finding and contend that it probably does not represent the true ability of the charter school to raise student achievement scores. Given this contention, it might have been worthwhile for the report to have presented a parallel analysis, omitting student achievement data in the first year of the charter.

[6] Miron, G., & Nelson, C. (2001) *Student academic achievement in charter schools: What we know and why we know so little*. Occasional Paper No. *41*. National Center for the Study of Privatization in Education, Teachers College, Columbia University. Retrieved June 19, 2009, from http://ncspe.org/keepout/papers/00041/590_OP41.pdf.

[7] Miron, G., Evergreen, S., & Urschel, J. (2008). *The impact of school choice reforms on student achievement*. Boulder and Tempe: Education and the Public Interest Center & Education Policy Research Unit. Retrieved November 20, 2008, from http://epicpolicy.org/files/CHOICE-10-Miron-FINAL-withapp22.pdf.

[8] A number of reviews and Web-based systems have facilitated the tracking of charter school research. Most noteworthy is an online searchable database of studies maintained by the National Charter School Research Project at the University of Washington. This useful database now contains 77 studies that consider charter school student achievement, although many lacked technical reports and, thus, might not be classified as actual research (see National Charter School Research Center. University of Washington. Retrieved June 18, 2009, from http://www.crpe.org/cs/crpe/view/projects/1?page=yes&id=3&parent.

[9] The Lubienski & Lubienski 2006 and the Braun, Jenkins, & Grigg 2006 studies both employed hierarchical linear models (HLMs) to examine the difference between charter schools and traditional public schools while taking into account differences in student and school characteristics. The Braun, et al. study found that charter schools were performing at levels noticeably lower than traditional public school in both reading and math at the fourth grade level. These differences were statistically significant. The Lubienski and Lubienski study looked at mathematic results in both grades 4 and 8. They found that charter schools performed at levels lower than traditional public school at grade 4—a finding that was statistically significant. At grade 8, the charter schools had slightly higher results, but this finding was too small and was not found to be statistically significant.

Lubienski, S. T., & Lubienski, C. (2006). School sector and academic achievement: A multilevel analysis of NAEP mathematics data. *American Educational Research Journal*. *43*(4), 651-698.

Braun, H., Jenkins, F., and Grigg, W. (2006). *A Closer Look at Charter Schools Using Hierarchical Linear Modeling* (NCES 2006-460). U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences. Washington, DC: U.S. Government Printing Office.

Hoxby, C.M. (2004 September). *A Straightforward Comparison of Charter Schools and Regular Public Schools in the United States.* Cambridge, MA: Department of Economics, Harvard University.

Nelson, F. H., Rosenberg, B., & Van Meter, N. (2004) Charter school achievement on the 2003 National Assessment of Educational Progress. Washington, DC.: American Federation of Teachers.

[10] The CREDO report indicated that their study covered more than 70% of the nation's charter schools. Our own calculations suggest that the study may have covered 65% of the nation's charter schools. More details are needed to know whether the CREDO estimate included Massachusetts and whether this estimate is based on all the charter schools in Colorado and Illinois or only the charter schools in Denver and Chicago.

[11] The findings in the figures (and discussed in the text) omit Massachusetts. Also, the researchers prepared separate press releases for each state, except Massachusetts.

[12] Loveless, T. (2003). The 2003 Brown Center report on American education: Charter schools: Achievement, accountability, and the role of expertise. Washington, DC: The Brookings Institution.

[13] Miron, G., Coryn, C., & Mackety, D. (2007). Evaluating the impact of charter schools on student achievement: A longitudinal look at the Great Lakes states. Boulder and Tempe: Education and the Public Interest Center & Education Policy Research Unit. Retrieved November 20, 2008, from http: http://epicpolicy.org/files/EPSL-0706-236-EPRU.pdf.

[14] Data for these studies are readily available from state assessment systems. Merging these datasets with school-level demographic data is relatively easy. The longitudinal design is employed when school- or group-level results are compared over two or more years. Studies in this category vary considerably depending on whether and how well they control for demographic differences between the charter and non-charter public schools being compared. Some studies use blocking of data and draw comparisons between subgroups of students based on ethnic background, FRL status, Title 1 funding status, special education status, etc. Other studies have controlled for differences using regression analyses, which take into account all of the same demographic characteristics.

[15] Solmon, L., Paark, K., & Garcia, D. (2001). *Does charter school attendance improve test scores?* Arizona: Goldwater Institute 's Center for Market-Based Education.

[16] Ballou, D., Teasley, B., & Zeidner, T., (2008). Charter schools in Idaho. In M. Berends, M.G. Springer, H.J. Walberg (Eds.), *Charter school outcomes.* New York, NY: Lawrence Erlbaum Associates.

Bilfulco, R., & Ladd, H.F. (2006). School choice, racial segregation and test-score gaps: Evidence from North Carolina's charter school program. *Journal of Policy Analysis and Management, 25*(1), 31-56.

Gronberg, T., & Jansen, D.W. (2005). Texas charter schools: An assessment in 2005. Austin: Texas Public Policy Foundation.

Miron, G., Cullen, A., Applegate, B., & Farrell, P. (2007). *Evaluation of the Delaware charter school reform: Final Report*. Dover, DE. Delaware State Board of Education. Retrieved June 18, 2009, from http://www.wmich.edu/evalctr/charter/de_cs-eval_final_report.pdf

Sass, T. R. (2006). Charter schools and student achievement in Florida. Gainesville, FL: American Education Finance Association.

Zimmer, R., Buddin, R, Chau, D., Gill, B., Guarino, C., Hamilton, L., Krop, C., McCaffrey, D., Sandler, M., & Brewer, D. (2003). *Charter school operation and performance: Evidence from California.* Santa Monica: RAND.

[17] Zimmer, R., Gill, B., Booker, K., Lavertu, S., Sass, T.R., & Witte, J. (2009) Charter schools in eight states: Effects on achievement, attainment, integration, and competition. RAND.. Retrieved June 21, 2009, from http://www.rand.org/pubs/monographs/2009/RAND_MG869.pdf.

[18] Matched student designs vary considerably in quality and scope. Although all use student-level data, the better studies include more schools, as well as more grades and subjects. Also, the better studies will examine changes over more years. The overall quality of these matched student designs also depends on the procedures and controls used to match students. Several quasi-experimental designs have developed and many of them use creative statistical methods to control for differences in the non-treatment group. For example, some (but not all) account for non-response error due to attrition of students or the fact that some students are held back and dropped from the analysis. Two increasingly common designs include propensity score matching and regression discontinuity.

[19] Abdulkadiroglu, A., Angrist, A., Cohodes, C., Dynarski, S., Fullerton, J., Kane, T., and Pathak, P. (2008). *Informing the debate: Comparing Boston's charter, pilot and traditional schools*. Boston: The Boston Foundation.

Hoxby, C.M., & Murarka, S. (2007). *Charter schools in New York City: Who enrolls and how they affect their students' achievement*. Cambridge, MA: National Bureau of Economic Research.

Hoxby, C.M., & Rockoff, J.E. (2004). *The impact of charter schools on student achievement*: A study of students who attend schools chartered by the Chicago Charter School Foundation. Cambridge, MA: Department of Economics, Harvard University.

[20] Using waiting lists for simulating random assignment is a promising idea, but a number of limitations need to be considered. As voucher research has shown, there is often considerable attrition in control groups constructed from waiting lists (see, e.g., Rouse, 1997). Moreover, fieldwork in charter schools conducted by Gary Miron suggests that charter school waiting lists are often insufficient for the construction of a randomized experiment. In many cases, such lists are out of date or contain an accumulation of names over a number of years. In the most extreme cases, these lists cannot be produced for review when requested and may not exist. Aside from questions about the validity of these waiting lists, it is nearly impossible to assess whether students on the lists had subsequently enrolled in other charter schools or had been exposed to other educational reforms. Some students who are accepted into a charter school may not take the place due to such requirements as the need for transportation or a requirement that parents volunteer at the school. Using the waiting lists to simulate random assignment first requires a review and audit of the waiting lists, and greater efforts must be made to identify and track students who do not assume a place at the charter school or who choose not to return to their traditional public schools.

[21] Based on our own work, we estimate that it would be possible to fund around 100 statewide matched student studies for the cost of this one large random assignment study.

[22] Gill, B., Timpane, P.M., Ross, K.E., & Brewer, D.J. (2001). *Rhetoric versus reality: What we know and what we need to know about vouchers and charter schools*. Santa Monica: RAND.

[23] Carnoy, M., Jacobsen, R., Mishel, L. & Rothstein, R. (2005). The charter school dust-up: Examining the evidence on enrollment and achievement. New York and Washington DC: Teachers College Press, and the Economic Policy Institute.

Hill, P. (2005). Assessing achievement in charter schools. In R. J. Lake (Ed.), *Hopes, fears, and reality: A balanced look at American charter schools in 2005.* Seattle: National Charter School Research Project, University of Washington.

National Alliance of Public Charter Schools. (2009). Charter school achievement: What we know, 5th Edition. April, 2009. Retrieved June 21, 2009, from http://tinyurl.com/oa4m7d

[24] Miron, G., Evergreen, S., & Urschel, J. (2008). *The impact of school choice reforms on student achievement*. Boulder and Tempe: Education and the Public Interest Center & Education Policy Research Unit. Retrieved June 21, 2009. from http://epicpolicy.org/files/CHOICE-10-Miron-FINAL-withapp22.pdf.

[25] For example, if a student scored 640 on the SAT (assuming a test mean of 500 and a SD of 100) that student's z-score would be 1.40. If a second student took the ACT and received a 27 composite (assuming a test mean of 20 and SD of 5) their z-score would also be 1.40. Although the z-scores are the same, few people would argue that each person would receive the same z-score if each took the other test. All the z-score transformation has done is place the student in a relative position to the "group." In this case, it is not even clear which group (within a state) was used to form the z-score transform.

Page 2 of the technical appendix includes a footnote that indicates the z-score transform may have pooled either grades or subject matter. We can see no defensible reason that the z-score transform be calculated from any pooled group – pooling over years, grades or subject tests. That is, it is only meaningful (given the two stated assumptions) if calculated within year, within grade, and within subject content. Moreover, unless the "group" distributions are approximately normal, the equal z-scores (1.40) actually mean very different things. Said another way, converting raw scores to z-scores does not change the shape of the group distribution; it just forces the mean to be 0.0 and the SD to be 1.0. So to believe that the conversion of different state achievement test scores to a z-score scale provides a common score scale for interpretation requires two major assumptions. First, the content of the different tests must have substantial overlap and this must be true across all grade levels of all assessments. Second, the distribution of scores for all assessments, at all grades, in all years must be approximately the same. Unfortunately, the former assumption would be difficult (although possible) to validate, and while the latter assumption could be easily empirically validated, no information was provided in the technical appendix regarding the distributions of scores to allow readers to do so.

Regarding the assumption of substantial content overlap, we know of no comprehensive research that has examined the content overlap of the different state achievement tests. Moreover, in many states the purpose of the test changes as a function of the assessment grade. For instance, in Michigan the MEAP, a basic skills test, is administered before high school. But in high school the Michigan Merit Exam is administered, and it is more of an aptitude test than a basic skills test.

For all these reasons, the validity of the standardizing method used in the CREDO report is less than perfect and may undermine its conclusions. Unfortunately, CREDO researchers merely dismiss these issues in the statement "minor differences may remain after these adjustments" (pg. 13).

[26] The choice of term "twin" CREDO used to describe the matched case for each TPS student is potentially problematic, since there is a great deal of research activity focused on the study of familial twins. Both research methods and statistical analyses are well-established for that twin research. In fact, proper statistical analyses in that area must account for the clustered nature of the twin data. Since the new study's use of the word "twin" does not refer to

that twin research (or use those methods), it would be helpful to explicitly tell readers that the use of the term is unique to that study.

[27] Miron, G., Cullen, A., Applegate, B., & Farrell, P. (2007). *Evaluation of the Delaware charter school reform: Final Report*. Dover, DE. Delaware State Board of Education. Retrieved June 18, 2009, from http://www.wmich.edu/evalctr/charter/de_cs-eval_final_report.pdf

[28] It is not completely clear how student mobility is addressed. It appears that even if the student attending a charter school remains in the same charter school they could have more than one virtual twin if the first virtual twin moves to a charter school.

[29] However, if the virtual school were substantially different from the feeder schools then the matching procedure has adjusted the characteristics of the virtual school to better match the charter school characteristics and the difference between the two (feeder schools and the virtual school) may represent some of the hypothesized selection bias. So it remains unclear how effective the matching procedure was in reducing the unknown but nonzero selection bias in the charter school students.

[30] For example, it is not clear if the CREDO researchers simply used an asymptotic covariance matrix of the estimates (thus accounting for heteroscedasticity present in the data) or took a more structured approach that would directly account for the clustering effects. We suspect the former since the clustering effect, while present in the virtual twin students, cannot be clearly represented by their school since they are without school, but the charter school students are with school. Moreover, clustering at the level of the State should have been better addressed in the CREDO analyses.

[31] There are several design approaches CREDO could have used with such a large database, including the use of the pre-test values with a regression discontinuity analysis. As stated earlier, a HLM might be most appropriate, and perhaps a matched-pairs design (which is used in true twins research) may have worked better. We suspect that matching on the pre-test values may have created some difficulties (hence the wide matching range), thus a propensity match could have been utilized or the researchers might even have used prior test scores as a covariate.

[32] Given the access to linked individual student records, we wonder why student growth modeling was not done, with school type (charter or traditional public) as a level-2 (or level-3) predictor depending on other covariates in the model. This type of analysis directly addresses what the CREDO researchers note to be the primary question of the evaluation: "…a current and comprehensible analysis about how well [charter schools] do educating their students" (pg. 1). If the focus of the study is on examining the "effects" of charter schools on student achievement then student achievement should be modeled within the nested structure of the data. This would appropriately account for the intra-class correlation present at each level of the analysis.

[33] For example, if reading and math were correlated at .7, this would mean that about 50% of the variance in reading is indistinguishable with math and vise-a-versa. In that case, a treatment that changes math scores would necessarily also affect about 50% of the variance in reading scores. It would be easy—and wrong—to then attribute the change in reading to be 'caused' by the treatment when, in fact, the change in reading scores reflects the shared variance with math scores. An unconfounded picture of reading (or math) can be obtained by including the non-dependent variable in the model as a predictor.

[34] Some states do not have caps on the numbers of charter schools but they do have restrictions on where charter schools can operate. The CREDO researchers did try to account to account for the manner in which the caps impacted the number of charter schools, but this attempt still could not capture the different ways in which caps are interpreted and eventually influence the numbers of charter schools.

[35] Miron, G., and Nelson, C. (2002). *What's Public About Charter Schools?: Lessons Learned About Accountability and Choice*. Thousand Oaks, CA: Corwin Press.

Miron, G. (2005). "Strong Charter School Laws Are Those That Result in Positive Outcomes." Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, April. Retrieved June 18, 2009, from http://www.wmich.edu/evalctr/charter/aera_2005_paper_charter_school_laws.pdf.

[36] Husén, T. (1990). "Strategy rules for education reform: An international perspective on the Spanish situation." In. T. Husén (Ed.) *Education and the global concern*. Oxford: Pergamon.

[37] Chi, W. C., & Welner, K. G. (2008). Charter Ranking Roulette: An Analysis of Reports that Grade States' Charter School Laws. *American Journal of Education*, *114*(2), pp. 273-298.

[38] Miron, G. (2005). "Strong Charter School Laws Are Those That Result in Positive Outcomes." Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, April. Retrieved June 18, 2009, from http://www.wmich.edu/evalctr/charter/aera_2005_paper_charter_school_laws.pdf.

The Think Tank Review Project is made possible by funding from the Great Lakes Center for Education Research and Practice