Measuring Charter School Effectiveness Across States

April 30, 2008

Kevin Booker
Duncan Chaplin
Eric Isenberg

Submitted to:
New Leaders for New Schools
30 West 26th Street
New York, NY 10010

Project Officer: Dianne Houghton

Submitted by:
Mathematica Policy Research, Inc.
600 Maryland Ave., SW, Suite 550
Washington, DC 20024-2512
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

Project Director: Duncan Chaplin
ACKNOWLEDGMENTS

The authors would like to thank the many people who contributed to this report, especially Chris Mathews, Allison Jack, Nick Stableski, Michael Gross, Dianne Houghton, Kurt Hyde, and Jonathan Schnur at New Leaders for New Schools for the important work they are doing and their support of our on-going efforts. We would also like to thank the many staff at the charter schools for all of their hard work compiling the data used in this report.

A number of staff at Mathematica Policy Research, Inc., assisted in this effort. Several researchers provided useful input, including Brian Gill, Steve Glazerman, and Aaron Douglas. Margaret Sullivan managed the collection of student-level data from the charter schools with assistance from Bosha Hoch. Tara Anderson collected the state-level data. Keith Rathbun and Joel Smith helped with programming. Amanda Bernhardt and Donna Dorsey provided editing, word processing, and production support.
THE MATHEMATICA POLICY RESEARCH
VALUE-ADDED MODEL

INTRODUCTION

New Leaders for New Schools, a non-profit organization committed to training school principals, heads the Effective Practices Incentive Community (EPIC), an initiative that offers financial awards to effective educators. New Leaders and its partner organizations have received more than $88 million from the U.S. Department of Education in 2006 and 2007 to support EPIC. Through this initiative, New Leaders will offer two types of financial awards to educators in four urban school districts and a consortium of charter schools: (1) a reward for principals and instructional staff in schools that are effective in raising student achievement, and (2) a financial incentive to document effective practices at award-winning schools. New Leaders will publicize their findings on effective practices online.

New Leaders contracted with Mathematica Policy Research, Inc. (MPR) to help design the methods for identifying effective schools and teachers. The approach used for each partner differs, depending on the priorities of the partner and the type of information available to measure school and teacher performance. This report presents the method used to identify effective schools for a consortium of 99 charter schools in 18 states and the District of Columbia during the first year of this project. MPR will work with New Leaders and the charter school consortium to revise the model in future years, and to incorporate any additional data that become available. The identification of effective teachers will be addressed in a later report.

METHOD FOR MEASURING SCHOOL EFFECTIVENESS

Many commonly used measures of school effectiveness, such as average test score levels or the percentage of students meeting the state proficiency standard, do not provide an accurate measure of school effectiveness because they are likely to be affected by students’ prior abilities and accumulated achievements and by current non-school factors like parents’ socio-economic status. Better measures of school effectiveness focus on how much a school contributes to test score improvements for its students. MPR follows this approach, basing its measures on student test score growth.
This technique, called a “value-added model” (VAM), has been used by a number of prominent researchers (Meyer 1996; Sanders 2000; McCaffrey et al. 2004; Raudenbush 2004; Hanushek et al. 2007). VAMs aim to measure students’ achievement growth from their own previous achievement levels. Many VAMs also control for student characteristics such as eligibility for free or reduced price lunch to account for factors that systematically affect the academic growth of different types of students. Thus, VAMs “handicap” both the students’ starting point and factors affecting their growth over the year. Because a value-added model accounts for initial student performance differences across schools, it allows schools with low baseline scores to be identified as high performers and vice versa.

A VAM provides a better measure of school effectiveness than relying on gains in the proportion of students achieving proficiency. Proficiency gains measure growth only for students who happen to cross the proficiency cut-point, but VAMs incorporate achievement gains for all students, regardless of their baseline achievement levels. In addition, unlike school-wide proficiency rates, which are affected by changes in the composition of students, VAMs track individual students over time. See Potamites and Chaplin (2008) for more details.

MPR uses a VAM to estimate the effect of each charter school on student performance in 2006-07, controlling for the prior performance of those students and a set of student demographic variables. Key aspects of the MPR model are outlined here, with a more detailed technical description in the appendix.

Data Requirements for Participation

Each charter school was asked to provide at least two years of data on test scores and student demographics for all tested students in all tested grades, except for students for whom baseline test scores were not available. For instance, in states that begin testing in third grade, elementary schools were not expected to provide past test scores for third graders. Neither were elementary or middle schools expected to provide test scores for students in the lowest grade served by the school. High schools, however, were expected to provide middle-school baseline scores (typically from 8th grade) for students who had attended a public school in the same state. All schools that provided data for at least 6 students’ current and past test scores were included in the model.

Of the 99 charter schools in the original charter consortium, 97 had the necessary data for inclusion in the model. Those schools represent 17 states and the District of Columbia, with 16 schools each from Florida and Illinois, 15 from California, 14 from the District of Columbia, 6 each from Indiana, Massachusetts, and New York, 3 each from Ohio and Texas, 2 each from Colorado, Missouri, and North Carolina, and 1 each from Arizona, Maryland, Mississippi, New Mexico, Pennsylvania, and Wisconsin. The only school in Tennessee had to be dropped because of a lack of the state-level data needed for the analysis, as explained in the appendix.
For analyzing the data and making awards, there are three grade ranges: elementary (grades 2-6), middle (grades 7-8), and high (grades 9-11). Of the 97 schools included in the analysis, 71 serve at least one elementary school grade, 53 serve at least one middle school grade, and 20 serve at least one high school grade. Forty-one schools span two grade ranges and three schools serve all three grade ranges. The ranking of any school within a grade range is based only on the students in that grade range. For example, a school serving grades 6 to 8 is ranked in the elementary school category based solely on its sixth graders, and in the middle school category based on its seventh and eighth graders.

Test Score Standardization

Because the VAM includes test scores for multiple grades, subjects, and years, as well as scores for different states that administer different exams, the scores must be standardized so that they fit comparable scales. MPR transforms the test scores by subtracting from each student’s score the state-wide mean for that subject, grade, and year, and dividing by the state-wide standard deviation for that state, subject, grade, and year. This yields a standardized score that equates each student to the average student in the state, and is comparable across schools within each state.

To allow comparison of test scores across different states, MPR adjusts student scores using state average scores and standard deviations from the National Assessment of Educational Progress (NAEP). Details of the adjustment method are given in the appendix.

The Value-Added Model

A student’s performance on a single test will be an imperfect measure of ability, so MPR employs a statistical technique known as “instrumental variable estimation” to obtain a more accurate measure of prior student achievement. The MPR model incorporates information on students’ performance on the test in the other subject from the prior year to measure prior student achievement. In other words, the measure of prior performance in math incorporates the measures of prior performance in English and vice versa.

The MPR VAM aims to measure how much a given school has raised student test scores after accounting for factors out of the school’s control. In addition to a student’s test score in that subject in the previous grade, the VAM includes a set of variables that statistically control for factors that can affect the academic growth of individual students: free or reduced price lunch status, limited English proficiency, special education status, gender, and ethnicity.

The MPR model accounts for the time that students who changed schools during the school year spent in their charter school. MPR allocates credit to a school based on the fraction of time the student spent at the school, known as the school “dosage.” Thus the model incorporates students who spent part of the year outside the charter school.
Precision of the School Rankings

MPR estimates the precision of the school performance measures. One way to illustrate the uncertainty associated with estimated school rankings is to examine the 90 percent confidence interval for each school's ranking. This gives the best and worst rankings for a school that fall within the margin of error associated with that school's estimated performance measure.

Figures 1, 2, and 3 show the confidence intervals for the school rankings in the elementary, middle, and high school grade ranges. The straight diagonal line is the ranking of each school in that grade range, with the best schools having the lowest rankings. The jagged line above it shows the high rank for schools; the jagged line below it shows the low rank for schools. Since the model is used to identify the best-performing schools, the region of interest is the lower left of the graph, documenting the precision of the top-ranked schools. For example, Figure 1 shows that, accounting for the uncertainty in the precision of school rankings, the top 10 percent of elementary schools, i.e., the top 7 schools, are all ranked at least 31st out of 71 schools. The results are somewhat less precise for middle schools, as the top 10 percent, (i.e., the top 5 schools), at worst rank 29 out of 53, somewhat below average. The results for high schools are similar to those for middle schools: the top 2 of 20 high schools rank only in the top 13 of 20. Further information on the precision of the estimates is contained in the appendix.

Model Limitations and Potential Extensions in Future Years

The MPR VAM has a number of limitations common to most VAMs. In future years, MPR plans to explore a number of extensions to address these limitations.

- The results are somewhat imprecise, especially for middle and high schools. In future years MPR will try to obtain data on a larger fraction of students in each school and perhaps on additional subjects and years to help improve the precision of the estimated school performance measures.

- Only math and English language arts scores for tested grades are covered in the analysis. This could create biased results for schools if the omitted grades and subjects differ systematically from those for which we have data. This could be a particular problem for the entering grade of a school, as there is some evidence that students often experience a drop in test scores during their first year in a charter (Booker et al. 2007). MPR may incorporate data on additional grades and subjects as they become available.

- Data are missing for a substantial fraction of students. MPR may modify the model to impute predicted values for students missing prior test scores but about whom there is enough other information to make an informative prediction about what their missing scores were likely to have been.
• School enrollment is reported for the school year, but the testing date varies across schools, so the dosage variables do not measure the time spent in a school from one test to another. MPR may modify the dosage measures to account for time spent between testing dates rather than between the beginning and end of the school year.

• The model may not control adequately for some variables that are not measured. One example is the extent to which more motivated parents systematically send their children to particular schools. Consequently, these schools may be given credit for test score gains that were caused by motivated parents. A second example is the extent to which students’ peers exert an influence on their test scores. MPR may modify the model to incorporate the possibility of peer effects associated with average characteristics of the students at the school.

• Smaller schools may be more likely to receive awards than larger schools in part because of greater random variation found in smaller samples of students—i.e., as a result of the luck of the draw in any particular year, rather than actual performance differences. MPR may perform additional tests to assess the possibility that small schools are more likely to be misclassified as highly effective and highly ineffective, a concern raised by Kane and Staiger (2002). If school size is related to ranking in a substantively important way resulting from random variation rather than true performance differences, MPR will explore alternative ways of dealing with this problem. One possibility would be using a “shrinkage estimator,” a statistical technique that “shrinks” the school effects toward the average, with greater shrinkage for small than for large schools. MPR also may estimate average performance across a number of years to improve the precision of measures of school effects.
REFERENCES


Appendix A

Technical Details of The Value-Added Model

Estimation Sample

Each charter school included in the final model provided MPR with test scores for their students in 2006-07, as well as for at least one prior year. The VAM is run separately for students in elementary school grades (2-6), middle school grades (7-8), and high school grades (9-11). Out of the 71 schools with elementary school grades, 39 also have middle school grades, including three schools that have students in all three grade-level groups. In addition to those three schools there are five other schools with students in both the middle school and high school grade groups.

Some students are excluded from each model due to insufficient data, which decreases the sample size. For the elementary school English language arts model, the initial dataset has a student sample of 9,258 students in grades 2-6, excluding grades where no students have the necessary data for inclusion in the model. Of those students, 157 are missing 2006-07 test scores, an additional 2,381 students are missing prior test scores either in that subject or in the other subject used as an instrumental variable, and 39 students are missing the gender or ethnicity variables.

After excluding these students, the elementary school estimation sample includes 6,681 2nd through 6th grade students at 71 schools, an average of 94 students per school. The sample size is similar for the math model, meaning that when both subjects are included elementary schools in our model have on average 188 observations on student test score growth.

For the middle school English language arts model, the initial dataset has a student sample of 6,270 students in grades 7-8, excluding grades where no students have the necessary data for inclusion. Of those students, 161 are missing 2006-07 test scores, an additional 1,376 students are missing prior test scores either in that subject or in the other
subject used as an instrumental variable, and one student is missing the gender or ethnicity variables. After excluding these students, the middle school estimation sample includes 4,732 7th through 8th grade students at 53 schools. This gives an average of 89 students per school. The sample size is similar for the math model, meaning that when both subjects are included middle schools in our model have an average 178 observations on student test score growth.

Many high schools do not test their students in every grade, often testing them only in 10th grade. For these schools, the model links each student’s 2006-07 high school test score back to the student’s most recent prior test score, usually an 8th grade score from 2004-05.

For the high school English language arts model, the initial dataset has a student sample of 1,342 students in grades 9-11, excluding grades where no students have the necessary data for inclusion in the model. Of those students, 63 are missing 2006-07 test scores and an additional 515 students are missing prior test scores either in that subject or in the other subject used as an instrumental variable. After excluding these students, the high school estimation sample includes 764 9th through 11th grade students at 20 schools. This gives an average of 38 students per school. The sample size is similar for the math model, meaning that when both subjects are included high schools in our model have an average 76 observations on student test score growth.

**Dosage Variables for Students Who Attended Multiple Schools during the 2006-07 School Year**

MPR uses data on the number of days each student was enrolled at a school to construct a school dosage variable that accounts for student mobility within the school year. Each dosage variable is equal to the percentage of the school year that the student spent at that school. For each student, the dosage variable at the school attended will be less than or equal to one; for all other schools, the dosage variable will be zero. Because a school is unlikely to be able to have an appreciable educational impact on a student who spends a very short time enrolled there, the dosage variable is set to zero for students who spent less than two weeks at a school and to one for students who spent all but two weeks or less at a school.

**Standardizing Test Scores**

To compare student performance across different states, grades, subjects, and years, MPR standardizes the test scores by subtracting the state-wide mean from each student’s score, and dividing each score by the state-wide standard deviation, where means and
standard deviations are calculated separately for each state, grade, subject, and year.\(^1\) MPR then uses data from the National Assessment of Educational Progress (NAEP) exam, a test given to a sample of students in each state, to adjust for differences in average student achievement across states.\(^2\) The process first estimates each state’s NAEP mean and standard deviation in each grade, subject, and year, using the 4th and 8th grade NAEP averages for that state in 2002-03, 2004-05, and 2006-07,\(^3\) then adjusts each student’s test score for the difference between the state NAEP mean and standard deviation and the national NAEP mean and standard deviation, creating a standardized score that is comparable across states, grades, and years.

**CONTROLLING FOR MEASUREMENT ERROR**

One of the key control variables in the VAM is the student’s prior year test score (for elementary school students this is a 2005-06 test). Any single test score contains measurement error, so including it as an explanatory variable can lead to attenuation bias in the estimate of the pretest coefficient and to bias of unknown direction in the other coefficients, including school dosage variables. To correct for this measurement error, the model uses two-stage least squares (2SLS) with the student’s prior test score in the other subject as an instrumental variable (IV) for the prior same-subject test score. In the model using elementary school grades and both subjects, the coefficient on the prior test score variable increases from 0.68 under ordinary least squares (OLS) to 0.89 with the IV. There is a similar increase if only one subject or if other grade levels are used. A Durbin-Wu-Hausman test for endogeneity strongly rejects the consistency of the OLS results for all variants of the sample, implying that 2SLS is the preferable model in this case.\(^4\)

---

\(^1\) Five states were unable to provide statewide means and standard deviations for 2006-07. For California and Texas MPR used the 2005-06 means and standard deviations to standardize the 2006-07 test scores based on evidence that these statistics changed very little over time. For Massachusetts and Maryland, these statistics were estimated based on data from the preceding three years. For Massachusetts, MPR estimated the missing mean and standard deviation from a predictive regression model that used data from 26 subject-by-grade Massachusetts test results from 2003-04 to 2005-06. For Maryland, the regression-based results were averaged with a method based on using the normal density function. Assuming that test scores are normally distributed, the mean and standard deviation are inferred from data provided on cut-points of the test score distribution and the percentage of students scoring above those cut-points. Tennessee was unable to provide sufficient data to estimate the means and standard deviations, so one charter in that state was omitted from the analysis.

\(^2\) Hoxby (2005), who writes “although NAEP is not a perfect bridge between states’ tests, it is by far the best available,” proposes using the NAEP in a way similar to the method described here to compare school performance across states.

\(^3\) To estimate state NAEP means and standard deviations for untested grades and years, MPR first linearly extrapolates the untested grades within the tested years, then extrapolates the untested years as a function of the means and standard deviations in the tested years. This method is justified based in part on the fact that while test performance varies substantially across states, the gaps between states are fairly stable across grades and years. Thus, models estimated without the NAEP adjustments produced results fairly similar to models with the NAEP adjustments with correlations in rankings of 0.95 or higher.

\(^4\) Davidson and McKinnon (1993) discuss this augmented regression method of testing for endogeneity. Hanushek et al. (2007) discuss twice-lagged test scores as instruments for the prior test score.
Unlike elementary and middle school students, high school students often have prior test scores that are more than one year prior to their 2006-07 test scores. Some states test math in one grade and English language arts in the next, so for those students the prior math score is from a different year than the prior English language arts score. MPR included those students in the model, allowing the prior test score in the other subject to be used as an IV even if it was not from the same school year.

**THE VALUE-ADDED MODEL**

The VAM equation used to estimate school impacts is

\[ Y_{i,j,t} = \beta_1 \hat{Y}_{i,j,t-1} + \beta_2 X_{i,t} + \beta_3 D_{i,t} + e_{i,j,t} \]

where, \( Y_{i,j,t} \) is the 2006-07 test score for student \( i \) in subject \( j \), \( \hat{Y}_{i,j,t-1} \) is the predicted prior test score for student \( i \) in subject \( j \), \( X_{i,t} \) is a vector of controls for individual student characteristics (described below), \( D_{i,t} \) is a vector of school dosage variables, and \( e_{i,j,t} \) is the error term. The value of \( \hat{Y}_{i,j,t-1} \) is assumed to capture all previous inputs into student achievement. The vector \( D_{i,t} \) includes one variable for each school in the model. Each variable equals the percentage of the year student \( i \) attended that school. The value of any element of \( D_{i,t} \) is zero if student \( i \) did not attend that school. The school performance measures are the coefficients on \( D_{i,t} \).

The model includes control variables for exogenous student characteristics. Ideally, these would include every factor outside of the school’s control so as to isolate the school effect on student achievement. In practice, however, the model can include only those variables in the model for which data are available. In addition to the student’s lagged test score, the school-by-grade dosage variables, and a constant term, the VAM regressions include the following student-level variables:

- Gender indicator
- Race/ethnicity indicators (white, African American, Hispanic, Asian, Native American)
- Free or reduced price lunch indicator
- Limited English proficiency indicator
- Special education status indicator

\[ \hat{Y}_{i,j,t-1} \text{ is assumed to capture all previous inputs into student achievement.} \]

\[ \text{The vector } D_{i,t} \text{ includes one variable for each school in the model. Each variable equals the percentage of the year student } i \text{ attended that school.} \]

\[ \text{The value of any element of } D_{i,t} \text{ is zero if student } i \text{ did not attend that school.} \]

\[ \text{The school performance measures are the coefficients on } D_{i,t}. \]

\[ Y_{i,j,t} = \beta_1 \hat{Y}_{i,j,t-1} + \beta_2 X_{i,t} + \beta_3 D_{i,t} + e_{i,j,t} \]

\[ \text{where, } Y_{i,j,t} \text{ is the 2006-07 test score for student } i \text{ in subject } j, \hat{Y}_{i,j,t-1} \text{ is the predicted prior test score for student } i \text{ in subject } j, X_{i,t} \text{ is a vector of controls for individual student characteristics (described below), } D_{i,t} \text{ is a vector of school dosage variables, and } e_{i,j,t} \text{ is the error term.} \]

\[ \text{The value of } \hat{Y}_{i,j,t-1} \text{ is assumed to capture all previous inputs into student achievement.} \]

\[ \text{The vector } D_{i,t} \text{ includes one variable for each school in the model. Each variable equals the percentage of the year student } i \text{ attended that school.} \]

\[ \text{The value of any element of } D_{i,t} \text{ is zero if student } i \text{ did not attend that school.} \]

\[ \text{The school performance measures are the coefficients on } D_{i,t}. \]

- Gender indicator
- Race/ethnicity indicators (white, African American, Hispanic, Asian, Native American)
- Free or reduced price lunch indicator
- Limited English proficiency indicator
- Special education status indicator

\[ ^5 \text{Missing student demographic variables are imputed using Stata’s “impute” command. This replaces missing data with the predicted value obtained from a regression of the missing variable on a set of other student demographic variables, using all observations in the data set with complete data.} \]
Because the VAM combines scores across multiple subjects and grades, most students will be included in the model twice, once for math and once for English language arts. The standard errors of the school performance measures are adjusted for the clustering of observations by student.

For each school and grade range, MPR calculates the standard error of the school’s estimated performance measure, which is primarily a function of the number of students with complete test score data in a given grade range (elementary, middle, or high school). Using these standard errors, MPR calculates that the school effects are jointly significant within each school type (elementary, middle, and high). MPR also uses these standard errors to calculate a 90 percent confidence interval for each school’s ranking, which corresponds to the ranking the school would have received if its school performance measure was at the high or low end of its 90 percent confidence interval. Figures 1, 2, and 3 show the confidence intervals for the rankings of schools in the elementary, middle, and high school categories.

An alternate way of describing the precision of the rankings is presented in Table A.1 below, which displays the ratios of the mean standard errors over the standard deviations of the value-added measures by grade level. This statistic estimates the fraction of the standard deviation of the school value-added measures due to noise. These statistics are much larger for charter schools than they are for Memphis public schools. This can be seen by comparing the results in Table A.1 with Table A.2 from Booker and Isenberg (2008). This occurs in part because the numbers of students per school and test score observations per student are lower for the charter schools than for the Memphis public schools.

### Table A.1. Mean Standard Error/Standard Deviation School Value-Added Measures

<table>
<thead>
<tr>
<th>School Type</th>
<th>Ratio</th>
<th># of Schools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>0.69</td>
<td>71</td>
</tr>
<tr>
<td>Middle</td>
<td>0.95</td>
<td>53</td>
</tr>
<tr>
<td>High</td>
<td>0.80</td>
<td>20</td>
</tr>
</tbody>
</table>

### Table A.2. Mean Standard Error/Standard Deviation School Value-Added Measures

<table>
<thead>
<tr>
<th>School Type</th>
<th>Ratio</th>
<th># of Schools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>0.38</td>
<td>107</td>
</tr>
<tr>
<td>Middle</td>
<td>0.36</td>
<td>30</td>
</tr>
<tr>
<td>High</td>
<td>0.22</td>
<td>30</td>
</tr>
</tbody>
</table>
Figure A.1. Elementary School Ranking Ranges

Source: Data collected by Mathematica Policy Research.

Note: The upper and lower lines are the upper and lower bounds of a 90 percent confidence interval around the school ranking, which is given as the middle line.
Figure A.2. Middle School Ranking Ranges

Source: Data collected by Mathematica Policy Research.

Note: The upper and lower lines are the upper and lower bounds of a 90 percent confidence interval around the school ranking, which is given as the middle line.
Figure A.3. High School Ranking Ranges

Source: Data collected by Mathematica Policy Research.

Note: The upper and lower lines are the upper and lower bounds of a 90 percent confidence interval around the school ranking, which is given as the middle line.