MPR Reference No.: 6325-530

MATHEMATICA
Policy Research, Inc.

Measuring School Effectiveness in Memphis

April 22, 2008

Kevin Booker
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 are grateful the many people who contributed to this report. Thanks go first to Chris Mathews, Jon Burchfield, Michael Gross, Dianne Houghton, Kurt Hyde, and Jonathan Schnur at New Leaders for New Schools for the important work they are doing and for their support of our on-going efforts. We would also like to thank John Barker, Commodore Primus, Tequilla Banks, and Kenna Jones at Memphis City Public schools for all of their hard work in putting together the data used in this report.

A number of staff at Mathematica Policy Research, Inc., assisted in this effort. Several researchers provided useful information relating to specific issues or studies, including Duncan Chaplin, Brian Gill, Steve Glazerman, and Jeffrey Max. Joel Smith helped with some preliminary analyses of the data. Daryl Hall and Donna Dorsey provided expert editing, word processing, and production support.

We are grateful for the assistance of all these people. Only the authors bear responsibility for the judgments and conclusions in this report.

CONTENTS

THE MATHEMATICA POLICY RESEARCH VALUE-ADDED MODEL

A. INTRODUCTION

New Leaders for New Schools, a non-profit organization committed to training high-quality urban public school principals, received several grants from the U.S. Department of Education in 2006 and 2007 to support the development of innovative teacher compensation strategies. New Leaders is partnering with four urban school districts and a consortium of charter schools to implement the Effective Practices Incentive Community (EPIC). Through this initiative, New Leaders will offer two types of financial awards to educators: (1) a reward for principals and instructional staff in schools that are effective in raising student achievement and (2) a financial incentive for teachers who are identified as effective and are willing to document and share their practices. As part of the EPIC project, New Leaders will collect information on the practices shared by these teachers and make them available to educators, locally and nationally, through an online system.

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 in the Memphis City Schools (MCS), one of the partner school districts, during the first year of this project. MPR will work with New Leaders and MCS to revise the model in future years, and to incorporate any additional data that become available. The results of this work were given to New Leaders but are not presented here in order to maintain the confidentiality of the individual schools. The identification of effective teachers will be addressed in later reports.

B. METHOD FOR MEASURING SCHOOL EFFECTIVENESS

Many commonly used measures of school effectiveness, such as average test score levels or the percentage of students who meet state proficiency standards, do not provide an accurate measure of school effectiveness because they are likely to be affected by students' prior ability and accumulated achievement, 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 their students. MPR follows this approach, basing its measures on student test score growth adjusted for factors that affect growth but are outside the school's control.

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, Kain, Rivkin, and Branch 2007). A VAM uses students' test scores from the previous year to measure how much they had learned prior to the current year and typically control for student characteristics, like eligibility for free or reduced price lunch, to account for factors that systematically affect the academic growth of different types of students. Since the model "handicaps" both the students' starting point and the factors that affect their growth over the year, a value-added measure of school effectiveness captures test score gains across schools holding all of these factors equal. 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.

MPR uses a VAM for Memphis to estimate the effect of schools on student performance in 2006-07, controlling for the prior performance of those students and a set of student demographic variables. MCS has provided test score data measuring student achievement over time, with Tennessee Comprehensive Achievement Program (TCAP) test scores available for grades 2-8 in math, English language arts, science, and social studies, and Gateway exam scores available for high school students in Algebra, English, and Biology. The TCAP and Gateway exams are the high-stakes exams for Tennessee. Key aspects of the MPR model are outlined here, with a more detailed technical description found in the appendix.

School Dosage

The MPR model differs from a typical VAM by accounting for the time that students who change schools during the school year spend in each school. In Memphis, these students compose 9.5 percent of the total. MPR allocates credit to a school based on the fraction of time the student spent at each school, which can be thought of as the school "dosage." Thus the model includes both students who attend multiple schools in a single year and even students who spent part of the year outside the district as long as they were enrolled in the MCS during testing in the prior and current years. Other researchers measuring school effectiveness omit many of these mobile students from their models, thereby ignoring important information about school effectiveness and potentially producing inaccurate results.

Test Score Standardization

Because the MPR model includes test scores for multiple grades, subjects, and years, the scores must be standardized so that they fit comparable scales. MPR transforms the test scores by subtracting from each student's score the district-wide mean for that subject, grade, and year, and dividing by the district-wide standard deviation for that subject, grade,

and year. This implies that the district average student test score in a given year equals zero, and that the average student test score "growth" from one year to the next is also set mechanically equal to zero.

A VAM provides a better measure of school effectiveness than relying on gains in the proportion of students achieving proficiency for two reasons. First, proficiency gains measure growth only for students who happen to cross the proficiency cut-point, but a VAM incorporate achievement gains for all students, regardless of their baseline achievement levels. Second, unlike school-wide proficiency rates, which are affected by changes in the composition of students, a VAM tracks individual students over time.

The Value-Added Model

The MPR VAM estimates a school's impact on student performance across all tested grades and subjects that the school serves. It aims to measure how much a given school has raised student test scores after accounting for factors out of the school's control. For each test score outcome, the VAM includes the student's corresponding test score in that subject in the previous grade and 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, ethnicity, whether the student switched schools between school years or within school years, and whether the student either skipped or was held back a grade. The model accounts for the fact that most students take the Gateway exam only once but in different grades.

Because a student's performance on a single test is an imperfect measure of ability, 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 from the prior year on students' performance on tests in other subjects to measure prior student achievement. For example, the measure of prior performance in math incorporates the measures of prior performance in English, science, and social studies.

Ranking Schools on Overall Performance

The VAM produces an estimated overall school effect across all grades and subjects the school serves. In addition to the overall school performance measure, MPR also estimates separate grade-by-subject school performance measures. Within each grade range (elementary, middle, or high school), the school scores are set equal to zero so that schools with a positive ranking are performing better than the average school included in the model, and schools with a negative ranking are performing worse than average. MPR also produces percentile rankings based on these scores, with the average school given a 50th percentile rank and other schools measured relative to this.

The rankings based on adjusted growth for one subject are sometimes quite different than those across all grades and subjects. Even the highest ranked schools may not excel in every grade and subject. For example, the top-ranked elementary school in the MCS based

on the overall VAM measure would rank 21st based on adjusted growth for math alone. This mediocre ranking is offset by higher growth in other subjects.

C. Model Limitations and Potential Extensions in Future Years

The MPR model has a number of limitations that are common to VAMs. In future years, MPR staff plan to explore a number of extensions to the model to address these limitations.

- ***Only tested grades and subjects are covered in the analysis.*** 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 who are 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 by MCS for the school year.*** Tests are not taken on the last day of each school year 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 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 a students' peers influence test scores. MPR may modify the model to incorporate the possibility of peer effects associated with average characteristics of the students at the school.

- ***The model assumes that all gains in the current year are attributed to the performance of the school in the current year.*** This ignores lingering effects of schools attended in earlier years. MPR may modify the model to allow for lingering effects of past schools.

- ***Smaller schools may be more likely to receive awards than larger schools.*** This can occur because of greater random variation found in smaller samples of students. MPR may perform additional tests for the possibility that small schools are more likely to be classified as highly effective and highly ineffective, a concern raised by Kane and Staiger (2002). If school size is related to ranking, MPR will explore the possibility of using a "shrinkage estimator," a statistical technique that "shrinks" the school effects toward the average, with greater shrinkage for small schools than large ones. MPR may also estimate average performance across a number of years to improve the precision of measures of school effects.

APPENDIX:

TECHNICAL DETAILS OF THE VALUE-ADDED MODEL

A. ESTIMATION SAMPLE

The MCS have provided MPR with TCAP test scores for students in grades 3-8 in 2002-03 and 2003-04, grades 2-8 in 2004-05 and 2005-06, grades 3-8 in 2006-07, and Gateway exam test scores for high schools students in 2002-03 through 2006-07. The main analysis is restricted to a group of schools identified by New Leaders, primarily schools with high percentages of students eligible for free or reduced price lunch. Some students are excluded from the model due to insufficient data, which decreases the sample size. The model excludes grade 2 students because there are no prior year test scores for them. For English language arts, the dataset has a student sample in 2006-07 of 53,592 students in grades 3-8. Of those, 4,211 are missing 2006-07 test scores and an additional 4,401 are missing prior test scores either in that subject or in another subject that is used as an instrumental variable; 33 have a grade change between 2005-06 and 2006-07 school years that was either negative or greater than two. Students who are held back or skip a grade have a lagged score that is standardized relative to the distribution of the grade they were in each year.

After excluding these students, there are 44,947 students at 137 schools in grades 3-8, an average of approximately 330 students per school. Finally, student test scores are excluded if they are in a school, grade, and subject in which five or fewer students had sufficient information for inclusion in the model. Only one student is excluded on this basis, giving a final estimation sample of 44,946 students. The sample size is similar for the other three tests. Schools with sufficient data on all four subjects—which includes almost all of the elementary and middle schools in this sample—have on average approximately 1,300 observations of student test score growth.

The average sample size is larger for middle schools than for elementary schools. On average, there are 200 students per grade in middle school but only 70 students per grade in

elementary schools, and the number of tested grades per school is similar. Elementary schools typically have three tested grades per school: 3, 4, and 5; a few include grade 6 as well. Most middle schools cover grades 6, 7, and 8. Both types of schools give tests in four subjects. Thus, on average there are about 840 test score observations per school for elementary schools and 2,400 test score observations per school for middle schools.

The model treats high schools slightly differently. Unlike TCAP tests in grades 2-8, Gateway exams are offered to students who have completed the corresponding course (algebra, English, or biology) and can be taken multiple times for students who fail the exam on their first try. The model uses the Gateway exam scores for students who take the exam for the first time in spring of 2007. When estimating separate subject-by-grade performance measures, the model treats grades 9-12 as a single grade.

In order to control for prior student performance, the model links each Gateway exam to the student's 8th grade TCAP exam in the corresponding subject (math, English language arts, or science). For students taking the Gateway English exam for the first time in 9th grade in 2006-07, their prior test score will be their 8th grade TCAP English language arts score from 2005-06, for 10th graders from 2004-05, 11th graders from 2003-04, and for 12th graders from 2002-03. Almost 20 percent of MCS students take the algebra Gateway exam in 8th grade. The model uses TCAP scores to measure performance for those students in 8th grade math. There is no measure of high school math performance for those students.

There are 7,701 students who took the algebra exam, 7,289 students who took the English exam, and 7,655 students who took the biology exam in grades 9-12 in spring of 2007. From the initial English sample of 7,289 students, 1,858 are missing their prior TCAP scores and 54 are missing one of the other variables in the model. After omitting these students, the final estimation sample for high school English is 5,377 students in grades 9-12 from 30 schools. No student is omitted on the basis of small samples within a school-by-subject-by-grade combination for TCAP English. The sample size and proportions are similar for the other two Gateway subjects. There are on average approximately 180 student test scores per subject per school. Most schools in the high school group have only students in grades 9 or above, but three schools also include grades 7 and 8 and have school performance measures based on both Gateway and TCAP scores. There are on average 700 test score observations per school. Thus the sample sizes used to measure high school performance are similar to those used for elementary schools but smaller than those used for middle schools.

B. Dosage Variables for Students Who Attended Multiple Schools During the 2006-07 School Year

MCS has provided administrative data tracking the percentage of the school year each student spent at every school he or she attended. MPR uses these data to account for student mobility within the school year by constructing school dosage variables for each school. These dosage variables are equal to the percentage of the school year that the student spent at that school. 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. To account for time spent at schools that are not part of the estimation sample, MPR constructs an additional dosage variable that equals the percentage of the school year that a student spent outside the schools in the estimation sample. This additional dosage variable combines time spent in Memphis schools that were not in the estimation sample and schools outside the city of Memphis. Thus the sum of the dosage variables for each student equals one hundred percent.

One concern of estimating a VAM is that schools may show gains in value-added test scores simply by exhorting new students to try harder on the test than they did in their old school. This effect could potentially be exacerbated by a dosage model because it includes students who switch schools mid-year. To test for this possibility, MPR estimated a model that excluded students who switched schools either within the school year or between school years. The results changed little, with correlations in school scores between the two models of 0.94, 0.88, and 0.95 for the three grade ranges (elementary, middle, and high schools.) Of the 17 top-scoring schools initially designated as award winners, only 2 would have dropped out of the top group had the model been estimated excluding students switching schools. The 18th top-scoring school was in its first year of operation and was therefore dropped from the analysis that excluded school switchers.

C. 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 the 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 average of the student's prior test scores in other subjects as an instrumental variable (IV) for the student's prior test score. The coefficient on the prior test score variable increases from 0.57 with no IV to 0.87 when other subject 2005-06 scores are used as an IV. A Durbin-Wu-Hausman test for endogeneity strongly rejects the consistency of the non-IV results, both for the overall model and for the separate subject-by-grade models, implying that 2SLS is the preferable model in this case.[1]

The 2SLS procedure is modified for the high school analysis, since not every student takes each Gateway exam for the first time in a given year and students in multiple grades are taking each exam in a given year. As a result, the 8th grade scores used in the high school model will be from different years, depending on the grade of the student in 2006-07.

D. THE VALUE-ADDED MODEL

The VAM equation used to estimate school impacts:

[1] 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.

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

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, including the composite "other school." 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_{it}, the elements of the vector β_3.

Because the overall VAM combines all subjects and grades, most students will be included in the model three to four times, once for each tested subject. The standard errors of the overall school performance measures are adjusted for the clustering of observations by student. This standard error is used to calculate a 90 percent confidence interval for each school. This confidence interval was used to report a high and low rank for each school, which correspond to the ranks the school would have received if their overall school performance measure was at the high or low end of their 90 percent confidence interval.

The ratios of the mean standard errors over the standard deviations of the value-added measures by school type are presented in Table A.1 below. This statistic estimates the fraction of the standard deviation of the school value-added measures that is due to noise.

Table A.1. Mean Standard Error/Standard Deviation School Value-Added Measures Memphis Public Schools

School Type	Ratio	# of Schools
Elementary	0.376	107
Middle	0.355	30
High	0.219	30

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

- Gender indicator
- Race/ethnicity indicators (white, African-American, Hispanic, Asian, Native American)

- Free or reduced-price lunch indicator
- Limited English proficiency indicator
- An indicator for special education status
- An indicator for switching schools between 2005-06 and 2006-07
- An indicator for switching schools within the 2006-07 school year
- An indicator for skipping a grade
- An indicator for being held back a grade
- Subject-by-grade indicators
- The combined percentage of the school year that the student spent outside schools in the estimation sample (which can also be thought of as an additional school-by-grade dosage variable)

When the model is run separately by grade and subject the same control variables are included, except for the subject-by-grade indicators. The separate high school models cover grades 9-12 and include indicators for student grade.

Skipping a grade and being held back a grade is under the control of schools to some extent. Consequently, it could be argued that they should not be included as control variables. However, we find evidence that in the MCS data the control variables do not affect school rankings. Value-added scores are closely related to average dosage-weighted student test score growth, which does not control for any student characteristics. For example, the correlation between the value-added measure and simple growth measure for 4th grade math is 0.98. For 7th grade math the correlation is 0.96.

E. ESTIMATING EFFECTS BY GRADE AND SUBJECT

The result of the VAM is a set of coefficient estimates capturing the effects of schools on student achievement across all the tested grades and subjects that the schools serve. MPR uses the model to produce separate overall effectiveness rankings for elementary, middle, and high schools. Elementary schools are defined as schools that only serve students in grade 6 or below, high schools as those that include students in grades 9 or above, and middle schools are defined as the remaining schools, including those that cover grades K-8. Each school's overall measure is based on the grades that they serve, so a K-8 school receives credit for the performance of their students in grades 3-5, even though they are included in the middle school category. Separate measures were also calculated by grade level and subject for comparison.

REFERENCES

Davidson, R., and J.G. MacKinnon. *Estimation and Inference in Econometrics*. 2nd ed. New York: Oxford University Press, 1993.

Hanushek, E.A., J.F. Kain, S.G. Rivkin, and G.F. Branch. "Charter School Quality and Parental Decision Making with School Choice." *Journal of Public Economics,* vol. 91, nos. 5-6, 2007, pp. 823-848.

Kane, Thomas J., and Douglas O. Staiger. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives,* vol. 16, no. 4, fall 2002, pp. 91-114.

McCaffrey, Daniel F., J.R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton.. "Models for Value-Added Modeling of Teacher Effects." *Journal of Educational and Behavioral Statistics,* vol. 29, no. 1, 2004, pp. 67-102.

Meyer, Robert H. "Value-Added Indicators of School Performance." In *Improving America's Schools: The Role of Incentives*, edited by Eric A. Hanushek and Dale W. Jorgenson. Washington, DC: National Academy Press, 1996

Raudenbush, S.W. "What Are Value-added Models Estimating and What Does This Imply for Statistical Practice?" *Journal of Educational and Behavioral Statistics*, vol. 29, no. 1, 2004, pp. 121-129.

Sanders, W.L. "Value-Added Assessment from Student Achievement Data—Opportunities and Hurdles." *Journal of Personnel Evaluation in Education*, vol. 14, no. 4, 2000, pp. 329-339.