Measuring School Effectiveness in Memphis—Year 2

Final Report

October 23, 2009

Liz Potamites
Duncan Chaplin
Eric Isenberg
Kevin Booker

MATHEMATICA
Policy Research, Inc.

Mathematica Reference Number:
6325-230

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

Submitted by:
Mathematica Policy Research
600 Maryland Avenue, SW
Suite 550
Washington, DC 20024-2512
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Duncan Chaplin

Measuring School Effectiveness in Memphis—Year 2

Final Report

October 23, 2009

Liz Potamites
Duncan Chaplin
Eric Isenberg
Kevin Booker

MATHEMATICA
Policy Research, Inc.

ACKNOWLEDGMENTS

The authors would like to thank the many people who contributed to this report, especially Chris Matthews, Cheryl Green, Dianne Houghton, Kurt Hyde, and Jonathan Schnur at New Leaders for New Schools. We would also like to thank John Barker, Julie Ballard, John Nickey, Jason Ogle, and Jeff Shive at Memphis City Public schools for their hard work in putting together the data used in this report and responding to our questions about the data.

Several staff at Mathematica Policy Research assisted in this effort, including Brian Gill and Steve Glazerman. Tara Anderson helped with the programming. Molly and Jim Cameron edited this report. Alfreda Holmes provided production support.

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

CONTENTS

THE MATHEMATICA POLICY RESEARCH VALUE-ADDED MODEL........................1

A. Introduction 1

B. Method for Measuring School Effectiveness... 1

1. Test Score Data 2
2. Test Score Standardization 3
3. School Dosage 3
4. The Value-Added Model.. 3
5. Ranking Schools on Overall Performance 4
6. Precision of School Rankings 4

REFERENCES...8

APPENDIX A: TECHNICAL DETAILS OF THE VALUE-ADDED MODEL A.1

TABLE

A.1 Reliabilities of the School Value-Added Measures A.4

FIGURES

1 90% Confidence Intervals for One-Year Full VAM Estimates, Elementary............. 5

2 90% Confidence Intervals for One-Year Full VAM Estimates, Middle School 6

3 90% Confidence Intervals for One-Year Full VAM Estimates, High School 7

THE MATHEMATICA POLICY RESEARCH VALUE-ADDED MODEL

A. Introduction

New Leaders for New Schools, a nonprofit 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 from the U.S. Department of Education tens of millions of dollars in financial support for EPIC. Through this initiative, New Leaders offers financial awards to educators in two urban school districts and a consortium of charter schools. Awards are meant to serve as a reward for principals and instructional staff in schools that are effective in raising student achievement and as a financial incentive to document effective practices at award-winning schools. New Leaders publicizes its findings on effective practices online.

New Leaders contracted with Mathematica Policy Research 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 second year of this project. Mathematica will work with New Leaders and MCS to revise the model in future years and to incorporate additional data that become available. The results of this work were given to New Leaders but are not presented here so as to maintain the confidentiality of the individual schools.

This year's model differs from last year's model in that we used a shrinkage estimator to help ensure that schools with small numbers of students in our model were not overrepresented at the top and bottom of the resulting performance measures. A shrinkage estimator is a statistical technique that "shrinks" the school effects toward the average, with greater shrinkage for schools whose results were less precisely estimated—typically smaller schools. More details on the shrinkage estimator can be found in the technical appendix.

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. This is because they are likely to be affected by students' prior ability and accumulated achievement, as well as by current non-school factors, such as parents' socioeconomic status. Better measures of school effectiveness focus on how much a school contributes to the test score improvements of its students. Mathematica follows this approach, basing its measures on student test score growth.

This technique, called a "value-added model" (VAM), has been used by several 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 account for both the students' starting point and the factors affecting their growth over the year. Because a value-added model accounts for initial student performance differences across schools, it allows schools having 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 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 the student population, VAMs track individual students over time. (See Potamites and Chaplin [2008] for more details.)

Ideally, VAMs estimate unbiased teacher and school effects. If students were randomly assigned to schools or classrooms and we had complete data on all students, our estimates would be unbiased. These conditions are unlikely. This means that our VAM estimates could be biased by unobserved factors that affect performance and are correlated with the schools or classrooms where a student is placed (Rothstein 2009). We control for prior test scores and observable characteristics in order to reduce the likelihood of such bias.[1] Kane and Staiger (2008) offer some evidence suggesting that unobservable student characteristics based on student assignment do not play a large role in determining VAM scores. Using data from the Los Angeles Unified School District, they compared (1) the difference in value-added measures between pairs of teachers based on a typical situation in which principals assign students to teachers and (2) the difference in student achievement between the teachers the following year, in which they taught classrooms that were formed by principals but then randomly assigned to the teachers. Kane and Staiger found that the differences between teachers' VAM scores before random assignment were a statistically significant and positive predictor of achievement differences when classrooms were assigned randomly. Because these results were gathered in schools in which the principal was willing to allow random assignment of classrooms to teachers, however, it is not clear if they generalize to other contexts.

Key aspects of the Mathematica model are outlined here, along with a more detailed technical description, found in the appendix.

1. Test Score Data

Mathematica uses a VAM for Memphis to estimate the effect of schools on student performance in 2007–08, controlling for the prior performance of those students. MCS has provided test score data measuring student achievement over time, with Tennessee Comprehensive Achievement Program (TCAP) test scores available for grades three through eight 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.

The model treats high school students slightly differently from other students because of differences in the tests. The elementary and middle school TCAP tests in grades three through eight are given once a year to each student. In contrast, the high school Gateway exams are offered to students who have completed the corresponding course (algebra, English, or biology) and can be

[1] Models are run both with and without other observable characteristics such as free and reduced price lunch status, English language learner status, special education status, gender, and ethnicity.

taken multiple times by students who fail on the first try. The model uses the Gateway exam scores for students who took an exam for the first time in spring of 2008. To control for prior student performance, the model links each Gateway exam to the student's eighth grade TCAP exam in the corresponding subject (math, English language arts, or science).

2. Test Score Standardization

Because the Mathematica model includes test scores for multiple grades, subjects, and years, the scores must be standardized so they fit comparable scales. Mathematica 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 these categories.[2] 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 also is set mechanically to zero.

3. School Dosage

The Mathematica model differs from a typical VAM by accounting for the time that students who change schools during the school year spend in each school. Students who spent time at more than one school in Memphis were seven percent of the analysis sample in 2007–08. Another six percent were at their school less than 90 percent of the time and were not in the district for the rest of the year.[3] Mathematica 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." Therefore, the model includes both students who attend multiple schools in a single year and 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.

4. The Value-Added Model

The Mathematica 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 controls for factors that can affect the academic growth of individual students: free or reduced price lunch status, limited English proficiency (LEP), special education status, grade level when tested, gender, ethnicity, whether the student switched schools between or within school years, and whether the student skipped a grade or was held back.

A version of the model was also run that included only the previous test score, not these other contextual factors. There are advantages and disadvantages to including these other variables. School rankings are very similar under either method (correlation of school rankings across one-year models

---

[2] Students who are held back or skip a grade have a lagged score that is standardized relative to the distribution of their grade level in each year.

[3] These are students who left the district and then returned between the baseline and follow-up tests. Students who left the district and did not return cannot be included because we do not have their end-of-year test scores.

was 0.996). NLNS used the model without other contextual factors to award schools and teachers in Year 2.

Because a student's performance on a single test is an imperfect measure of ability, Mathematica employs a statistical technique known as "instrumental variable estimation" to obtain a more accurate measure of prior student achievement. The Mathematica 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.

5. 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, Mathematica also estimates separate school performance measures for individual subjects.[4] Even the highest ranked schools may not excel in every grade and subject. For example, both the elementary and middle school that are top-ranked based on the overall VAM measure would rank sixth based on their VAM scores for science alone.

6. Precision of School Rankings

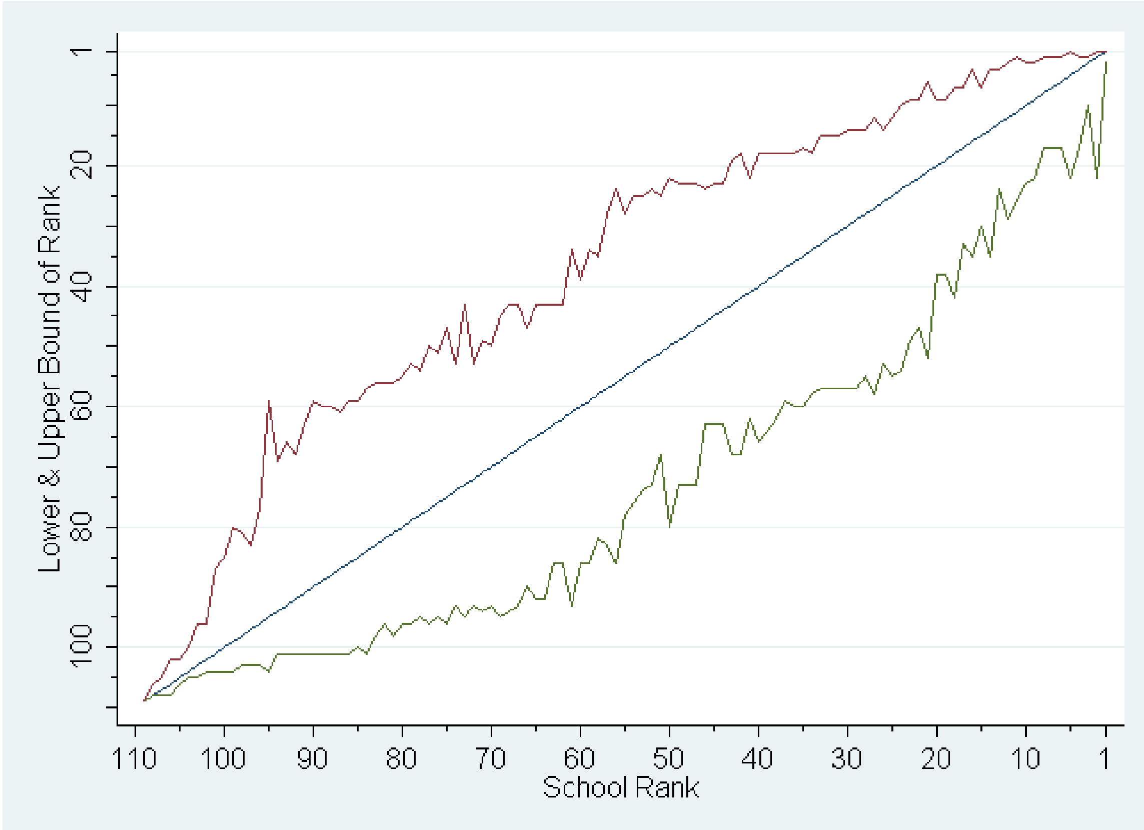
Mathematica estimates the precision of the school performance measures. One way to illustrate the uncertainty associated with estimated school rankings is to examine the 90 percent confidence interval for each school's ranking. This gives a school's best and worst rankings that fall within the margin 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. These rankings are based on the full VAM, using one year of performance data and one year of baseline data. Schools are judged on their performance in the 2007–08 school year. 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 the diagonal shows the best rank in each school's 90 percent confidence interval; the jagged line below the diagonal shows the worst rank for each school's confidence interval.

Since the model is used to identify the best-performing schools, the region of interest is the top right of the graph, documenting the precision of the rankings of the top-ranked schools. For example, Figure 1 shows that, given the uncertainty in our estimates of school rankings, with 90 percent confidence, the top 10 percent of elementary schools—that is, the top 11 schools—all are ranked no worse than the top 24 percent of schools (that is, 26th out of 109 schools). The results are slightly more precise for middle schools, as the top 10 percent—that is, the top 4 schools—rank no worse than the top 18 percent of schools (that is, 7th out of 40 schools, as shown in Figure 2). The results for high schools are the least precise: the top 10 percent of high schools (the top 4) rank no worse than the top 39 percent (that is, 14th of 36 schools) with 90 percent confidence (Figure 3).

[4] Within each grade range (elementary, middle, or high school), the school scores are set equal to zero so schools with a positive score are performing better than the average school included in the model, and schools with a negative score are shown as performing worse than average.

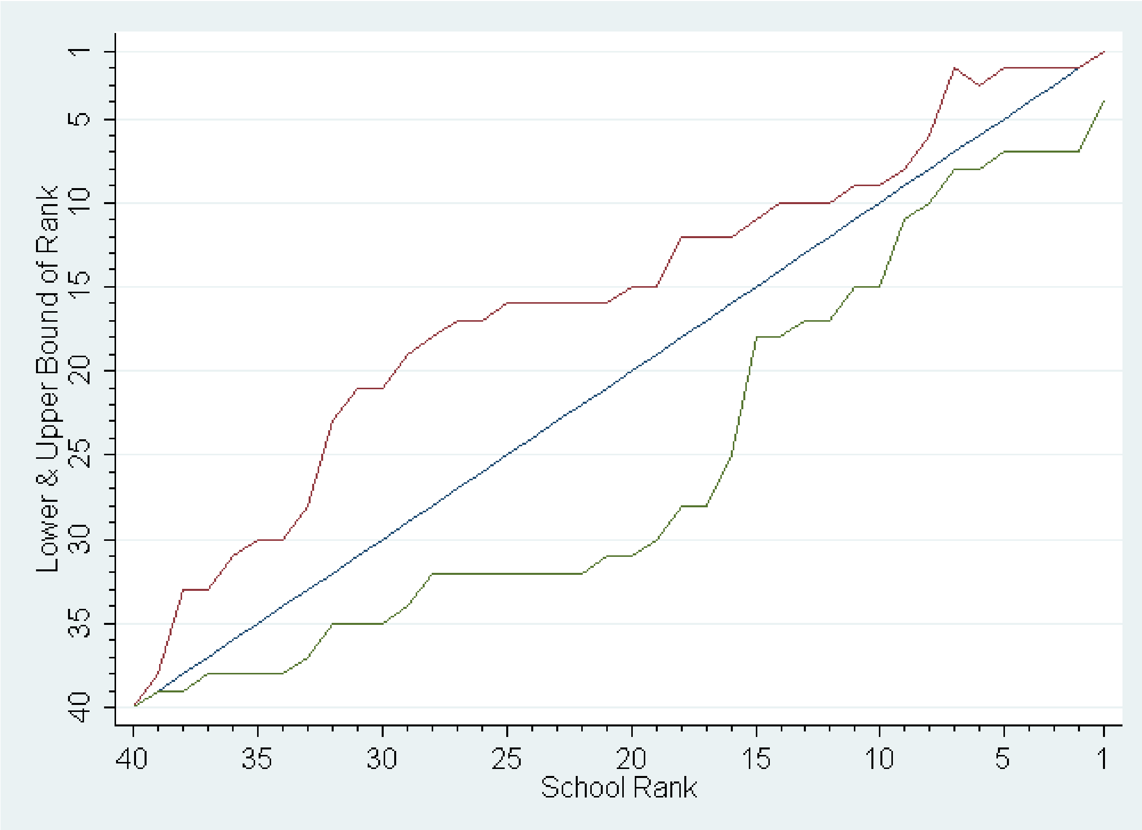
Figure 1. 90% Confidence Intervals for One-Year Full VAM Estimates, Elementary



Source: Data collected and analyzed 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.

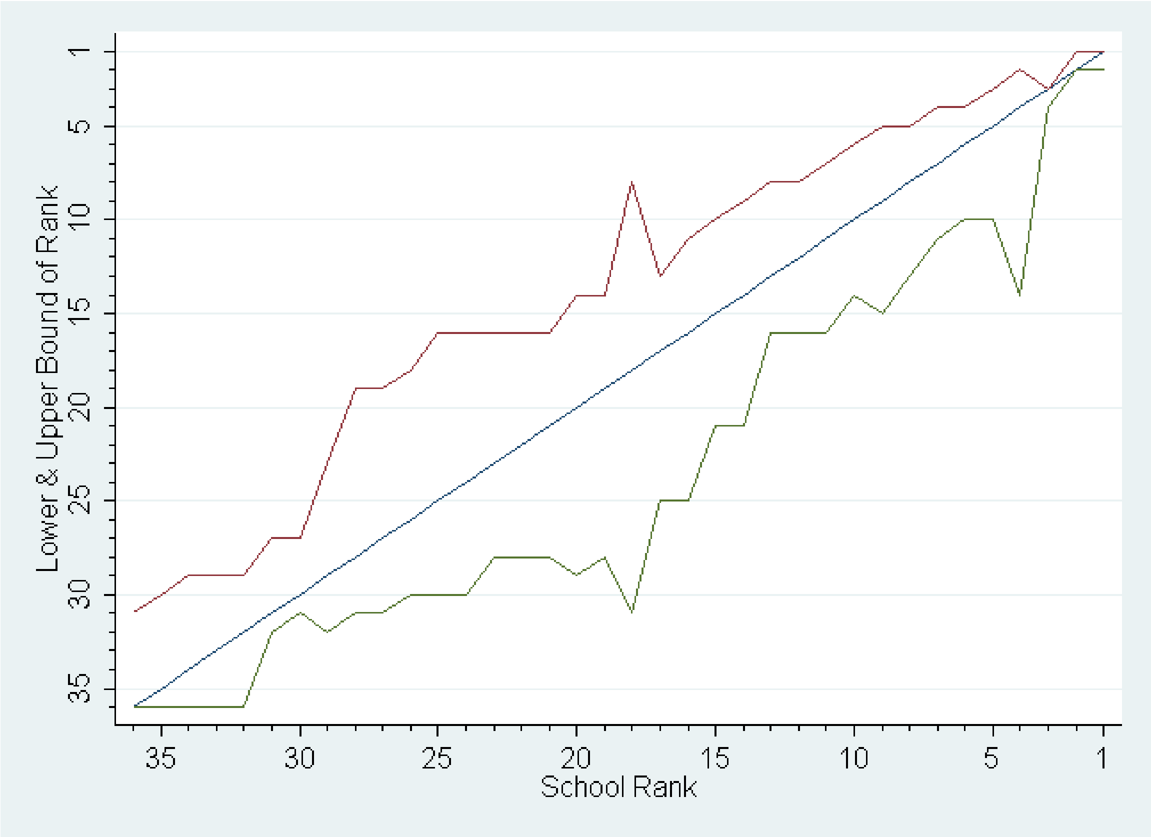
Figure 2. 90% Confidence Intervals for One-Year Full VAM Estimates, Middle School



Source: Data collected and analyzed 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 3. 90% Confidence Intervals for One-Year Full VAM Estimates, High School



Source: Data collected and analyzed 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.

REFERENCES

Ballou, Dale. "Value-Added Assessment: Controlling for Context with Misspecified Models." Working paper. Nashville, TN: Peabody College, Vanderbilt University, July 2007.

Booker, Kevin, and Eric Isenberg. "Measuring School Effectiveness in Memphis." Washington, DC: Mathematica Policy Research, 2008.

Carlin, Bradley P., and Thomas A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Boca Raton, FL: Chapman & Hall, 1996.

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, J.M. Markman, and S.G. Rivkin. "Does Peer Ability Affect Student Achievement?" *Journal of Applied Econometrics,* vol. 18, 2003, pp. 527-544.

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.

Huber, Peter J. "The Behavior of Maximum Likelihood Estimation Under Nonstandard Conditions." Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1, edited by L. M. LeCam and J. Neyman. Berkeley, CA: University of California Press, 1967.

Kane, Thomas J., and Douglas O. Staiger. "Estimating Teacher Impacts on Student Achievement, An Experimental Evaluation." Working paper no. 14607. Cambridge, MA: National Bureau of Economic Research, 2008.

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.

Morris, Carl N. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of American Statistical Association*, vol. 78, no. 381, 1983, pp. 47-55.

Potamites, Elizabeth, and Duncan Chaplin, "Ranking DC's Public Schools Based on Their Improvement in Math from 2006-2007." Washington, DC: Mathematica Policy Research, 2008.

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.

Rothstein, Jesse. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." Working paper no. 14442. Cambridge, MA: National Bureau of Economic Research, 2009.

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.

White, H. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica,* vol. 48, 1980, pp. 817–830.

APPENDIX A: TECHNICAL DETAILS OF THE VALUE-ADDED MODEL

A. Estimation Sample

The MCS have provided Mathematica with TCAP test scores for students in grades three through eight and Gateway exam test scores for high school students. Some students are excluded from the model due to insufficient data—most often a missing baseline test score. The model excludes grade three students because there are no prior year test scores for them (unless they failed third grade in the previous year). After excluding these students, there remained 48,114 tested students in 2007–08 who were matched with a prior test score. The main analysis was not restricted to schools eligible for awards but included all schools in Memphis with at least a total of 40 matched cases across both years and all subjects. In total, 109 elementary schools, 40 middle schools, and 36 high schools were included.

For students taking the Gateway English exam for the first time in 9th grade in 2007–08, their prior test score will be their 8th grade TCAP English language arts score from 2006–07, for 10th graders their score from 2005–06, for 11th graders their score from 2004–05, and for 12th graders their score from 2002–03. The model controls for the grade in which the student took the current test.

B. Dosage Variables for Students Who Attended Multiple Schools

MCS has provided administrative data tracking the percentage of the school year each student spent at every school he or she attended. Mathematica 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 fraction of the school year that the student spent at that school. Because a school is unlikely 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 there.

C. Controlling for Measurement Error

One of the key control variables in the VAM is the student's prior year test score—for elementary and middle school students, the 2006–07 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 0.88 when the average other subject prior scores are used as an IV.

D. The Value-Added Model

The VAM equation used to estimate school impacts:

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

where, $Y_{i,j,t}$ is the 2007-08 test score for student *i* in subject *j*, $Y_{i,j,t-1}$ is the predicted value for the prior test score for student *i* in subject *j*, $X_{i,t}$ is a vector of controls for individual student characteristics (including a constant and other variables described below), $D_{i,t}$ is a vector of school dosage variables, and $e_{i,j,t}$ is the error term. The value of $Y_{i,j,t-1}$ 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 $D_{i,t}$ of the elements of the vector β_3. The VAM is run jointly on all schools (elementary, middle, and high).The model includes control variables for exogenous student characteristics ($X_{i,t}$). These are chosen as factors outside of the school's control so as to isolate the school effect on student achievement. 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[5]
- Race/ethnicity indicators (white, African American, Hispanic, Asian, Native American)
- Free or reduced price lunch indicator[6]
- Limited English proficiency indicator[7]
- Special education status indicator
- First year at new school indicator
- Indicators for skipping a grade or failing a grade since the last test
- Indicator variables for grade level and subject and their interaction terms

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 (Huber 1967, White 1980). 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 estimated 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. Figures 1, 2, and 3 show the confidence intervals for the rankings of schools in the elementary, middle, and high school categories for the full one-year model.

E. Shrinkage Estimator

A new addition to the model this year is the shrinkage estimator, an empirical Bayes procedure outlined in Morris (1983). It uses an iterative process to jointly estimate the shrinkage weights for each school and the grand mean toward which all schools are shrunk. The basic idea is that

[5] Gender also was interacted with subject.

[6] Free or reduced price lunch also was interacted with grade.

[7] LEP also was interacted with subject.

imprecisely measured school estimates can be improved by "borrowing strength" from the overall estimate because the mean effect is more precisely estimated than any of the individual school effects. Each school effect is shrunk toward the overall mean using a weighted average of its individual estimate and the overall weighted mean (which is unknown). The first step is to form weights equal to the inverse of the square of the standard error plus an estimate of the variance of the overall mean (also unknown, we used the raw variance of the school effects as the starting point). Using these weights recalculate the overall mean and variance of the school effects and then calculate new weights and repeat until the process converges. Since each school's weight depends on the standard error of its original estimate, the Bayesian estimates for schools that are less precisely measured (with higher standard errors) will place more weight on the overall mean, compared to schools with lower standard errors.

The mean weight placed on the overall mean from the full one-year VAM estimates is 0.08 (the minimum weight across schools is 0.04 and the maximum is 0.46). The correlation between the standard error of each school's estimate and the weight placed on the overall mean is 0.98. The correlation between the estimates before and after shrinking is 0.99.

All estimated models included a constant term. After shrinking the estimates, the coefficients were also mean-centered. An alternative model would omit the constant and include the omitted schools category as a control variable.[8] The difference between the value-added estimate of any individual school relative to another is the same regardless of these modeling choices.

F. Precision of the VAM Estimates

Alternative ways of comparing the precision and variation of the rankings are presented in Table A.1 below. These include the mean standard error, the mean standard error squared, the standard deviation of the estimates, the ratio of the mean standard error to the standard deviation, and the reliabilities of the VAM estimates. The ratio of the mean standard error to the standard deviation can be interpreted as the fraction of the standard deviation due to noise. Reliabilities are a way of measuring the signal-to-noise ratio and are calculated as one minus the mean of squared standard error of the estimates, divided by the variance of the estimates.

Table A.1 shows that using the shrinkage estimator decreases the mean standard errors of the estimates but also slightly decreases the reliability of the estimates because the standard deviation of the estimates decreases by a greater proportion, since all estimates are shrinking toward a common mean.

[8] The omitted dosage variable is $1—\Sigma_s D_s$, where D_s s are the dosage variables for the included schools. There are many students in the model for whom we have information on their schools for only part of the year. Thus, the omitted schools are the ones those students had for the remaining time.

Table A.1 Reliabilities of the School Value-Added Measures

Model	Mean Standard Error (SE)	Mean Squared Std. Error (SSE)	Std. Dev. of Estimates (σ)	Mean SE/σ	Reliability (1-SSE/σ^2)
Elementary schools, N=109					
1 yr, full, not shrunk	.044	.002	.135	.330	.888
1 yr, full, shrunk	.043	.002	.123	.348	.876
Middle schools, N=40					
1 yr, full, not shrunk	.036	.001	.196	.184	.964
1 yr, full, shrunk	.035	.001	.172	.204	.956
High schools, N=36					
1 yr, full, not shrunk	.061	.004	.217	.282	.908
1 yr, full, shrunk	.056	.003	.173	.322	.887
All schools, N=185[10]					
1 yr, full, not shrunk	.046	.002	.169	.271	.918
1 yr, full, shrunk	.045	.002	.144	.302	.902

G. Robustness of Model to Alternative Specifications[11]

- **Peer effects.** Mathematica tested versions of the VAM that incorporate peer effects but do not use these control variables because the estimates of the peer effects were not robust to minor changes in specifications, and the effect on the school dosage variables was small. Our models which incorporated peer effects used two years of performance data and assumed constant school quality across years for each school. The estimated effects of peers were based on variation over time in the characteristics of the other students in the same school and grade. We calculated the mean and standard deviation of the test scores of other students and the percentage of other students who received free or reduced price lunch and included these variables in the VAM.[12] Similar to Ballou (2007), we find that the coefficients on peer effects were unstable to small variations in the model. In addition, adding peer effects had little effect on the relative rankings of schools. The school VAM estimates without peer effects were correlated at 0.99 or higher when compared to models that included (1) means and standard deviations of once-lagged peer test scores, (2) means and standard deviations of twice-lagged peer test scores (as recommended by Hanushek et al. 2003), or (3) the percent of students who received free or reduced price lunch.

---

[10] MCS charter schools and other schools that were ineligible for awards were included in the analysis for comparison sake.

[11] All of the results described in this section used models similar to the model described in Booker and Isenberg (2008).

[12] These means and percentages are dosage weighted in two ways. First, the peer effect variable for each school is weighted by the dosage for each student who attended that school. Second, the peer effect variable for each student is a dosage weighted average of the peer effect variables for each of the schools he or she attended.

- **Imputed missing baseline scores.** We ran a test to see whether or not imputing missing baseline test scores would improve our estimates. To do this test we assumed that mobile students for whom we did have data on baseline scores were similar to students who were missing baseline scores. We tested various models that imputed baseline test scores for these mobile students and found no method of imputation that would improve our models compared to dropping these students. To make comparisons we calculated the correlations between the results based on each of the imputation methods and the "true" model that included the actual pretest scores of mobile students. We also estimated a model excluding the mobile students. This is equivalent to assuming that they had the same value added as other students at those schools. This latter model had the highest correlation with the "true" model. Consequently, we chose not to impute missing baseline test scores.

- **Calculation of school enrollment.** Since tests are not taken on the last day of each school year, the dosage variables do not measure the time spent in a school from one test to another. To test whether this affected our measures of school performance, we compared our model, which measured dosage using the entire school year during which the tests were taken, to a model that measured dosage between the first day of the school year and the test date. We found that adjusting the dosage measures in this way did not lead to any substantial change in the VA estimates.