Measuring School and Teacher Effectiveness in the EPIC Charter School Consortium— Year 2

Final Report

October 23, 2009

Liz Potamites
Kevin Booker
Duncan Chaplin
Eric Isenberg

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 and Teacher Effectiveness in the EPIC Charter School Consortium—Year 2

Final Report

October 23, 2009

Liz Potamites
Kevin Booker
Duncan Chaplin
Eric Isenberg

MATHEMATICA
Policy Research, Inc.

ACKNOWLEDGMENTS

The authors would like to thank the many people who contributed to this report, especially Chris Mathews, Allison Jack, Jen Murck, Dianne Houghton, Kurt Hyde, and Jonathan Schnur at New Leaders for New Schools. 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.

Several staff at Mathematica Policy Research assisted in this effort. Researchers who provided useful input include Brian Gill, Steve Glazerman, and Aaron Douglas. Margaret Sullivan managed the collection of student-level data from the charter schools, with assistance from Virginia Knechtel and Claire Smither. Tara Anderson collected the state-level data. Ama Takyi helped with the preliminary programming and Joel Smith worked on the data from the District of Columbia. Jeremy Page did the programming that set up the data for teacher estimates. Molly and Jim Cameron edited this report. Alfreda Holmes provided production support.

CONTENTS

TABLES

FIGURES

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 to support 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 both a reward for principals and instructional staff in schools that are effective in raising student achievement and 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 methods used to identify effective schools and teachers for a consortium of over 140 charter schools in 17 states and the District of Columbia during the second year of this project. Mathematica will work with New Leaders and the charter school consortium to revise the model in future years and incorporate additional data that become available.

This year's model differs from that of last year in four respects. First, we used a shrinkage method to help ensure that schools (and teachers) with small numbers of students in our model were not overrepresented at the top (and bottom) of the resulting performance measures. Second, we added teacher-level estimates for charter schools. Third, we estimated models using two years of performance data rather than just one.[1] Finally, we changed the treatment of schools that cover multiple grade spans. Last year, those schools had multiple chances for winning awards (one for each grade span).[2] This year, we altered our model so each school had only one chance of getting an award. More details on these changes are presented below in the discussion of this year's methods.

B. Method for Measuring School and Teacher Effectiveness

Many commonly used measures of student outcomes aggregated at the classroom or school level, such as average test score levels or the percentage of students meeting the state proficiency standards, do not provide an accurate measure of the effectiveness of schools or teachers. This is because they are likely to be affected by students' prior abilities and accumulated achievements, as well as such other factors as parents' socioeconomic status. Better measures of effectiveness focus

[1] Last year we estimated models using only one year of performance (that is, one cohort of students in each grade level). This year we estimated models using both one and two years of a performance data. NLNS considered both models for the schools. The one- and two-year estimates for the schools were correlated at 0.94. NLNS gave out school awards based on the one-year models. The teacher awards were given out based on the two-year models. Using two-year models is particularly important for teachers because teacher estimates are generally less precise than school estimates due to the small sample sizes of students. Also teacher estimates are not stable over time (Sass 2008).

[2] Although schools could participate in more than one grade span category, that is, both the elementary and middle school categories for K through eighth grade schools, schools ultimately could receive no more than one monetary award, even if they were top-ranked in more than one category.

on how much a school or teacher 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 based on their own previous achievement levels. Many VAMs also control for such student characteristics 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 and teachers having students with low baseline scores to be identified as high performers and vice versa.

A VAM provides a better measure of 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 examine the achievement growth of 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.[3] 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.

Mathematica uses a VAM to estimate the effect of each charter school and teacher on student performance, controlling for the prior performance of those students. Key aspects of the Mathematica model are outlined here; a more detailed technical description is in the appendix.

1. Data Requirements for Participation

In order to estimate models covering two years of performance, each charter school was asked to provide data on at least two cohorts of students. For each cohort we needed math and reading

[3] Models were 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.

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 their third graders. Neither elementary schools nor middle schools were expected to provide baseline 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 eighth grade) for their students who had attended a public school in the same state if they were from a state that did not test in multiple grades in high school.[4] All schools that provided data on current and past test scores for at least 15 students were included in the model.

There were 145 charter schools in the second year EPIC consortium with the necessary data for inclusion in the model.[5] Those schools represent 17 states and the District of Columbia; 30 schools from California, 22 from the District of Columbia, 17 from Florida, 12 from Illinois, 10 from Massachusetts, 8 from Pennsylvania, 7 each from Colorado and New York, 6 each from Michigan and Texas, 5 each from Georgia and Indiana, 3 each from Louisiana and Ohio, and 1 each from Hawaii, Minnesota, Missouri, and New Mexico.

As mentioned in the introduction, schools that covered multiple grade spans were eligible last year for at most one award but could win that award based on their performance in any of the grade spans served. This year, we changed our methods so that each school competed in no more than one grade span. Schools were classified according to the majority of students served. Schools were ranked in the high school category if at least 50 percent of their students in the model were in grades 9 to 12. Middle schools were defined as schools that were not high schools but had at least 50 percent of their students in the model in grades 7 to 12. The elementary school category included all other schools. Furthermore, rather than running separate models for each grade span, all schools were included in the same model (with indicators for the grade level); the estimated coefficients then were ranked within each grade span. Of the 145 schools included in the analysis, 83 were considered to be elementary schools, 37 to be middle schools, and 25 to be high schools. Of the 83 schools with elementary school grades, 47 also have students in 7th grade and above, including one school with students in 9th grade or above. Of the 25 high schools, only one had students in 7th or 8th grade as well.

The teacher model included 908 teachers from the 114 schools with sufficient data available to estimate teacher effects. Sufficient data meant having at least 15 students with student-teacher links, end-of-year test scores and baseline test scores. Teachers were classified based on the grades of the courses they taught and could be eligible for awards based on their performance averaged across multiple spans if they taught multiple courses in different grades. There were 572 teachers included in the elementary grade range, 233 in the middle school grades, and 103 in the high school grades.

2. Test Score Standardization

Because the VAM includes test scores for multiple grades, subjects, and years, as well as scores from different states that administer different exams, the scores must be standardized so they fit

[4] There were only three high schools included in our models from states which tested students in multiple grades. These high schools were not required to obtain pre-high school test score data.

[5] One charter school submitted data and was included in our analyses but requested that they not be considered for an award.

comparable scales. Mathematica transforms the test scores by subtracting from each student's score the statewide mean for that test, subject, grade, and year, and dividing by the statewide standard deviation for those categories. This yields a standardized score that equates each student to the average student in the state and that is comparable across schools within each state.

To allow comparison of test scores across different states, Mathematica 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.

3. The Value-Added Model

A student's performance on a single test is an imperfect measure of ability, so Mathematica employs a statistical technique known as "instrumental variable estimation" to obtain a more accurate measure of prior student achievement. By instrumenting for the prior math score with the prior reading score and vice versa, the Mathematica model incorporates information on students' performance on the tests in both subjects in the prior year to measure prior student achievement.[6]

The Mathematica VAM aims to measure how much a given school or teacher has raised student test scores, after accounting for factors out of the school's control. In addition to a student's test score in a particular subject in the previous grade, the full 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. As mentioned previously, a version of the model was also run that included only the previous test score, not the other contextual factors. There are advantages and disadvantages to including these other variables. School rankings are very similar under either method (the correlation of school rankings in the one year models was 0.988). NLNS used the model without other contextual factors to award schools and teachers in Year 2.

Dosage Variables Constructed. The Mathematica model also accounts for the enrolled time of students who changed schools or teachers during the school year. It 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 attended the charter school for only part of the year. Dosage works similarly in the teacher model, allocating credit to a teacher based on the fraction of the time the student spent in that teacher's classroom. A student can contribute to estimates for multiple schools/teachers in a single year if he or she switches schools/classrooms during the year.

Shrinkage Estimator Used. Some VAMs implicitly and inadvertently favor smaller schools because of greater random variation found in smaller samples of students—that is, as a result of the luck of the draw in any particular year rather than actual performance differences (Kane and Staiger 2002). Similarly, teachers having fewer classes or smaller class sizes also may have a greater random variation due to their smaller student samples. One way of dealing with this phenomenon is to use a "shrinkage estimator," a statistical technique that "shrinks" the school or teacher effects toward an average of a larger group of schools or teachers, with greater shrinkage for schools or teachers whose results were less precisely estimated—typically those with fewer students. This year, Mathematica implemented a shrinkage estimator for schools and teachers. (Details can be found in the appendix.)

[6] Charter schools were asked to report test scores in two subjects: math and reading.

Imputed Missing Data for Demographics Only. Data are missing for a substantial fraction of students.[7] Mathematica imputed missing demographic data using methods explained in the appendix. We also considered imputing for students whose prior test scores were missing when there was enough other information to make an informative prediction about what their missing scores were likely to have been. In the summer of 2008, we ran models to test this method and determined that the imputed values of the missing test scores might not be sufficiently good to improve our models. More details are presented in "Measuring School Effectiveness in Memphis–Year 2" (Potamites et al. 2009).

Peer Effects Were Not Included in Model. The model may not control adequately for some variables that are not measured. One example is the extent to which students' peers exert an influence on their test scores. Mathematica considered modifying the model to incorporate the possibility of peer effects associated with the average characteristics of the students at the school. We determined that the best methods currently available for making these adjustments did not appear to be very robust. In addition, they did not appear to have an effect on the relative rankings of schools. (More details are presented in Potamites et al. 2009.)

C. Precision of the Rankings

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

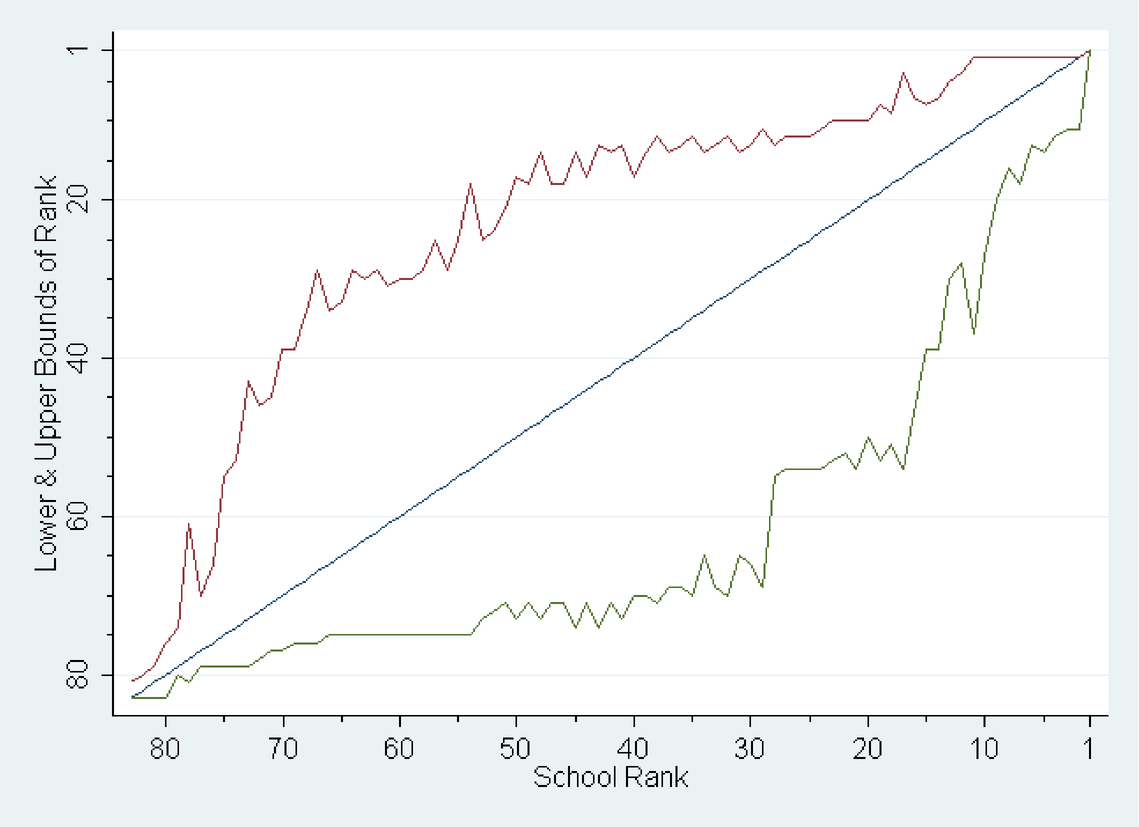
Figures 1, 2, and 3 show the confidence intervals for the school rankings in the elementary, middle, and high school grade ranges. These rankings are based on the full VAM, using one year of performance data. Schools are judged here on their performance in the 2007–08 school year.[8] The straight diagonal line is the ranking of each school, 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 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 (n=8) are all ranked in at least the top 22 percent (18th out of 83 schools). The results shown in Figure 2 are only slightly less precise for middle schools, as the top 10 percent (n=4) rank in the top 27 percent at worst (10th out of 37). The results for high schools shown in Figure 3 are somewhat less precise at the top end of the distribution than those for the middle schools: the top 10 percent of high schools (n=3) rank only in the top 32 percent (8th of 25).

[7] Ethnicity, gender, and special education status were missing for less than 1 percent of the final one-year analysis sample. Free or reduced price lunch status was missing for 6 percent and limited English proficiency status was missing for 12 percent.

[8] Except for the 11 schools from Indiana and Michigan, where students are tested in the fall of the school year, these schools are judged by their performance in 2006–07.

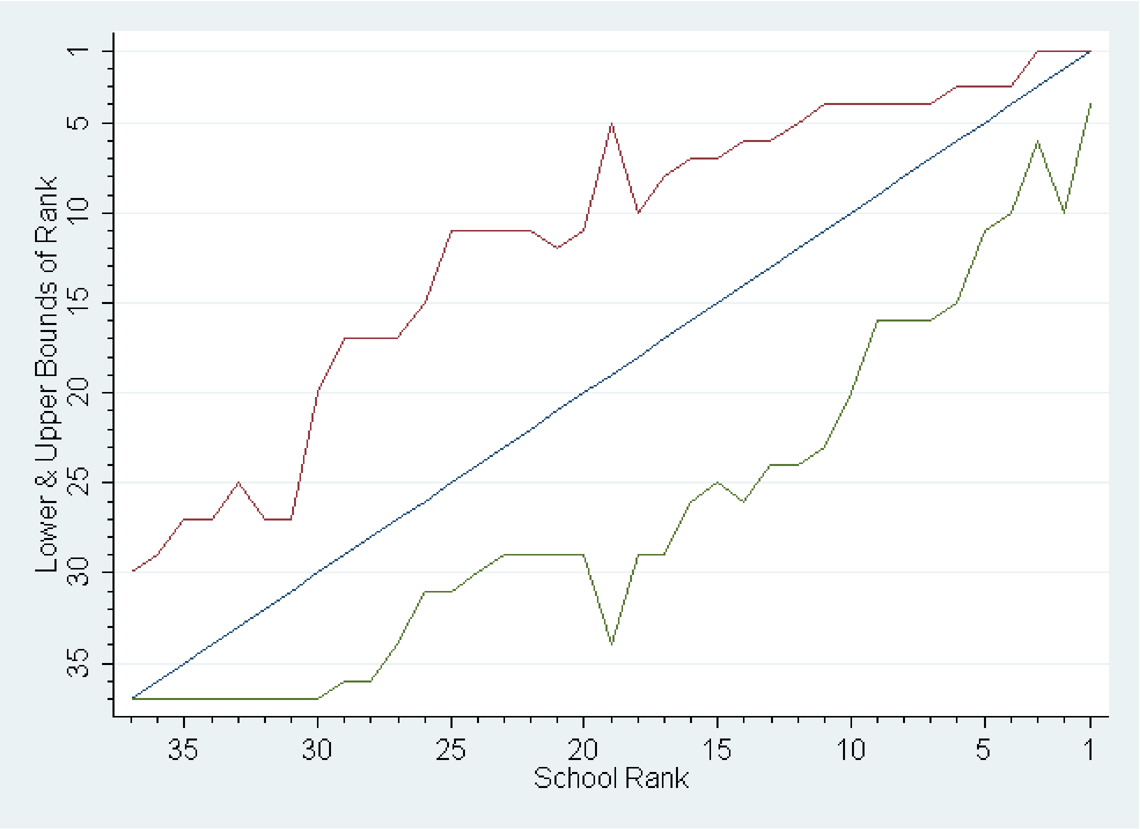
Figure 1. 90% Confidence Intervals for One-Year Full VAM Estimates, Elementary



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.

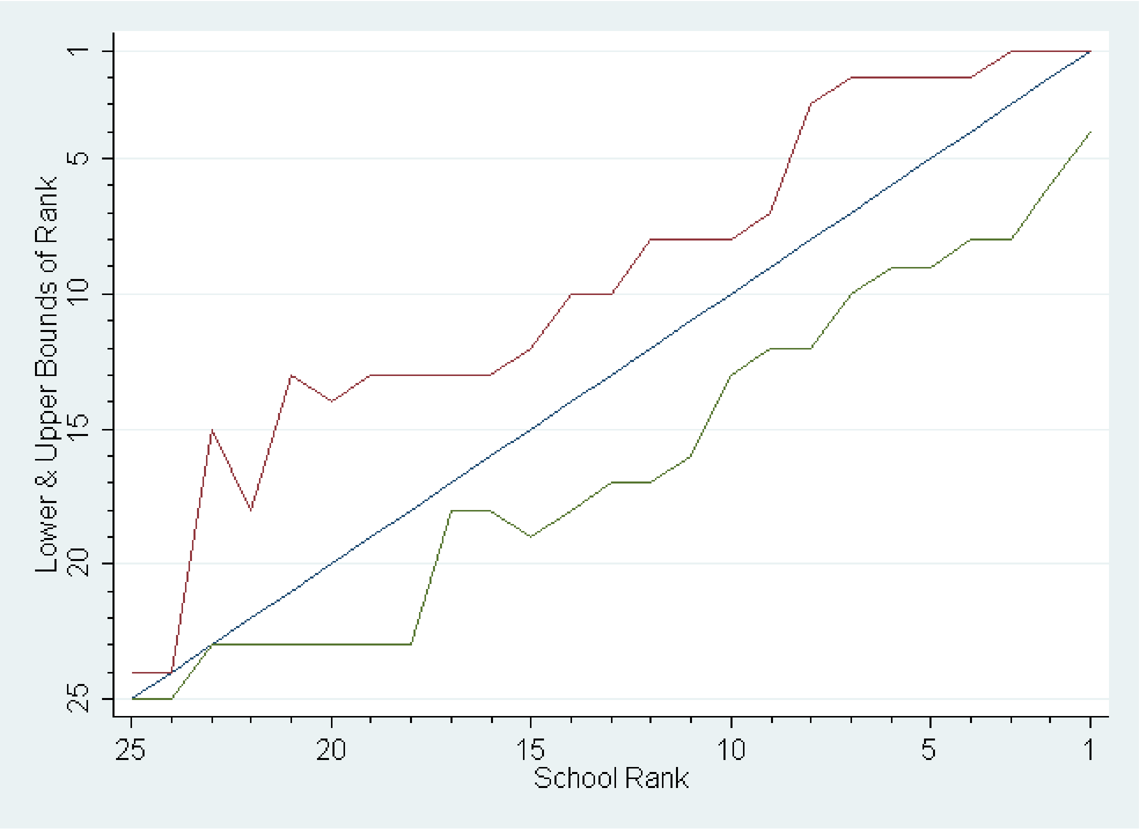
Figure 2. 90% Confidence Intervals for One-Year Full VAM Estimates, Middle Schools



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 Schools



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.

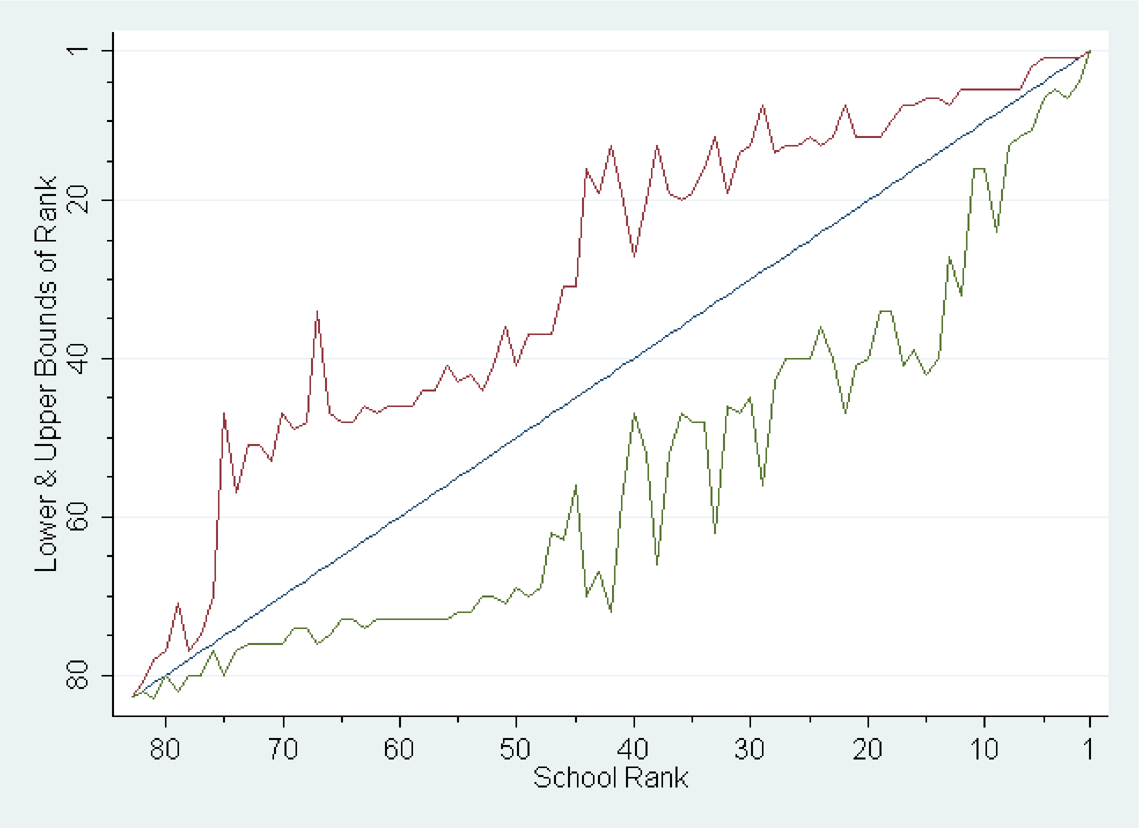
One way to improve the precision of the estimates is to use additional years of data. Mathematica estimated school VAM models based on two years of performance and found that the estimated standard errors decreased from an average of .081 for the one-year full model after shrinking to .051 for the two-year full model after shrinking.[9] Similarly, estimated standard errors decreased in the teacher model, from an average of .160 for the one-year full model after shrinking to .132 for the two-year full model following shrinking. Further information on the precision of the estimates is contained in the appendix.

Figures 4, 5, and 6 show the 90 percent confidence intervals for the school rankings where two years of performance data were available. For schools with available data (116 out of 145), the estimates are based on their performance in both the 2007–08 and the 2006–07 school years, except that schools in Michigan and Indiana were judged based on 2005–06 and 2006–07. The 29 schools with performance results from 2007–08 only are also included in the two-year results. As expected,

[9] For comparison, the average standard error in the one-year full model before shrinking was .088. So the shrinkage estimator reduced standard errors by 8 percent, while adding another year of data further reduced the mean standard error by 37 percent. See Table A.1 in the appendix.

the precision increases. With 90 percent confidence, the top 10 percent of elementary schools all are ranked in at least in the top 12 percent (10 out of 83), the top 10 percent of middle schools all are still within the top 20 percent (7 out of 35), and the top 10 percent of high schools are still in the top 24 percent (6 out of 25). The top schools in each category are mostly the same in both the one- and two-year models. There is one switch among the elementary schools and one among the middle schools.

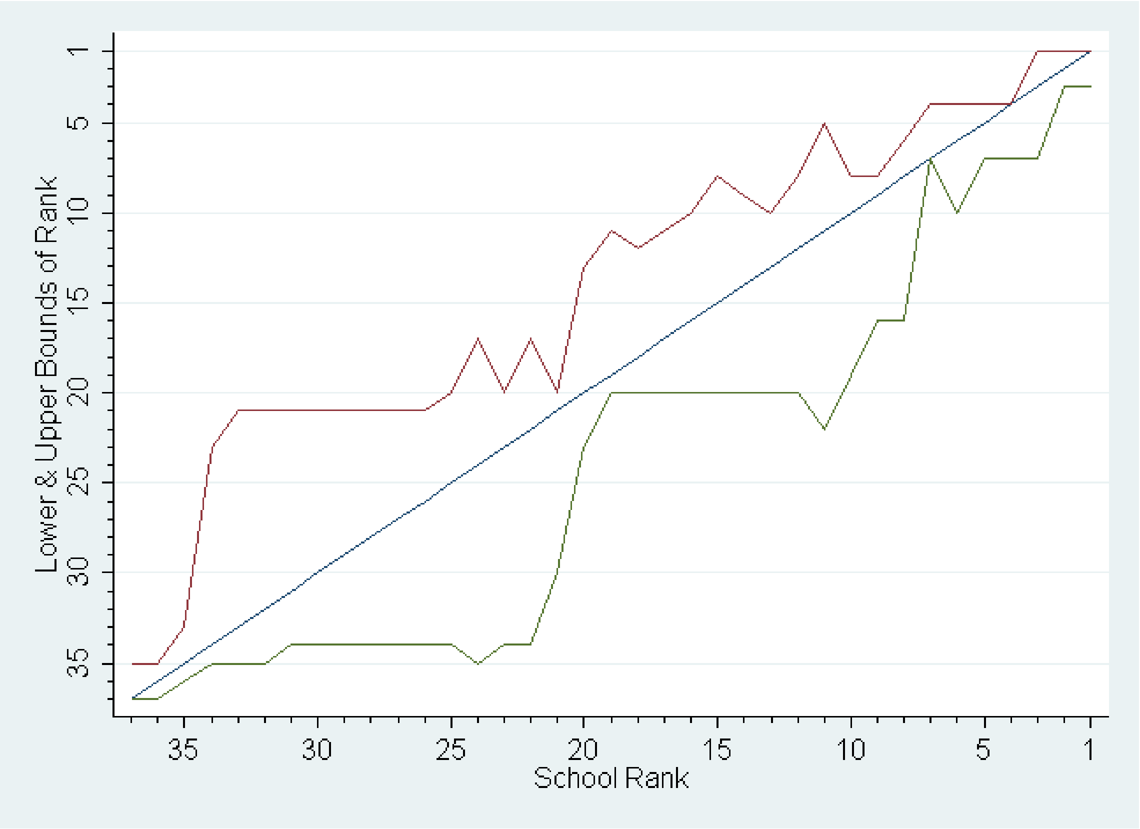
Figure 4. 90% Confidence Intervals for Two-Year Full VAM Estimates, Elementary



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.

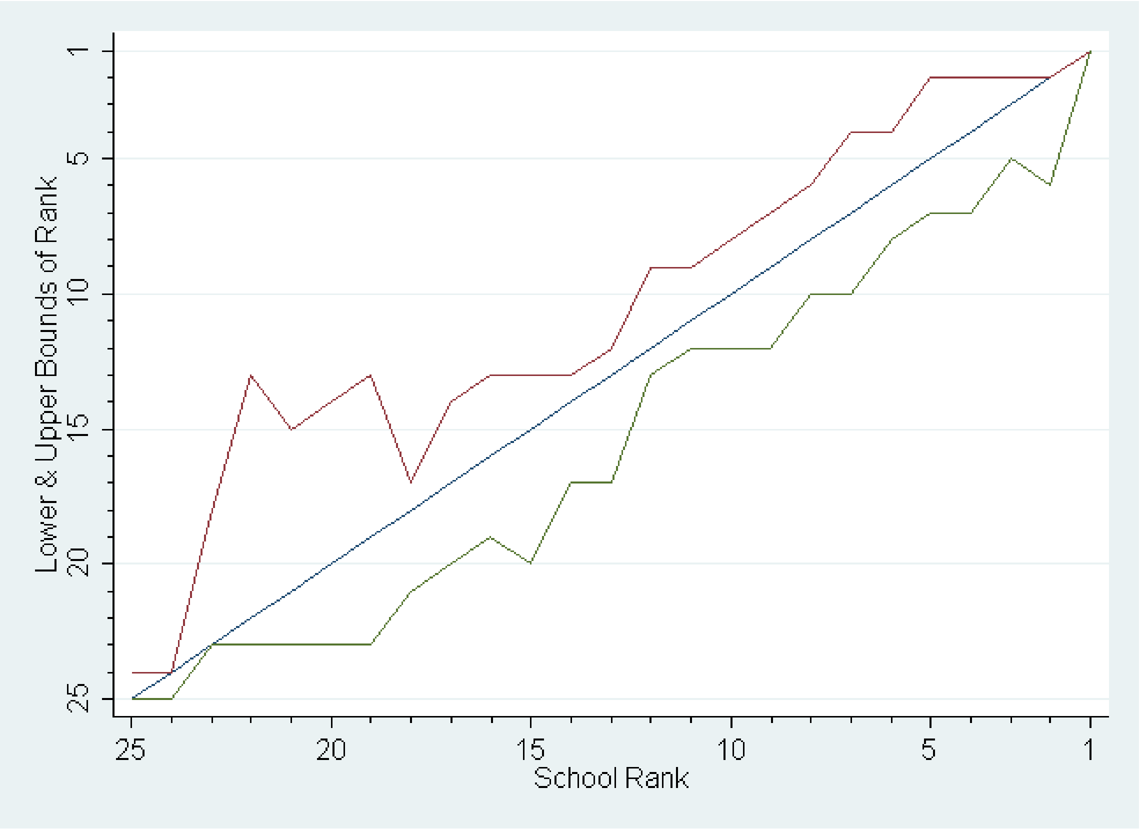
Figure 5. 90% Confidence Intervals for Two-Year Full VAM Estimates, Middle Schools



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 6. 90% Confidence Intervals for Two-Year Full VAM Estimates, High Schools



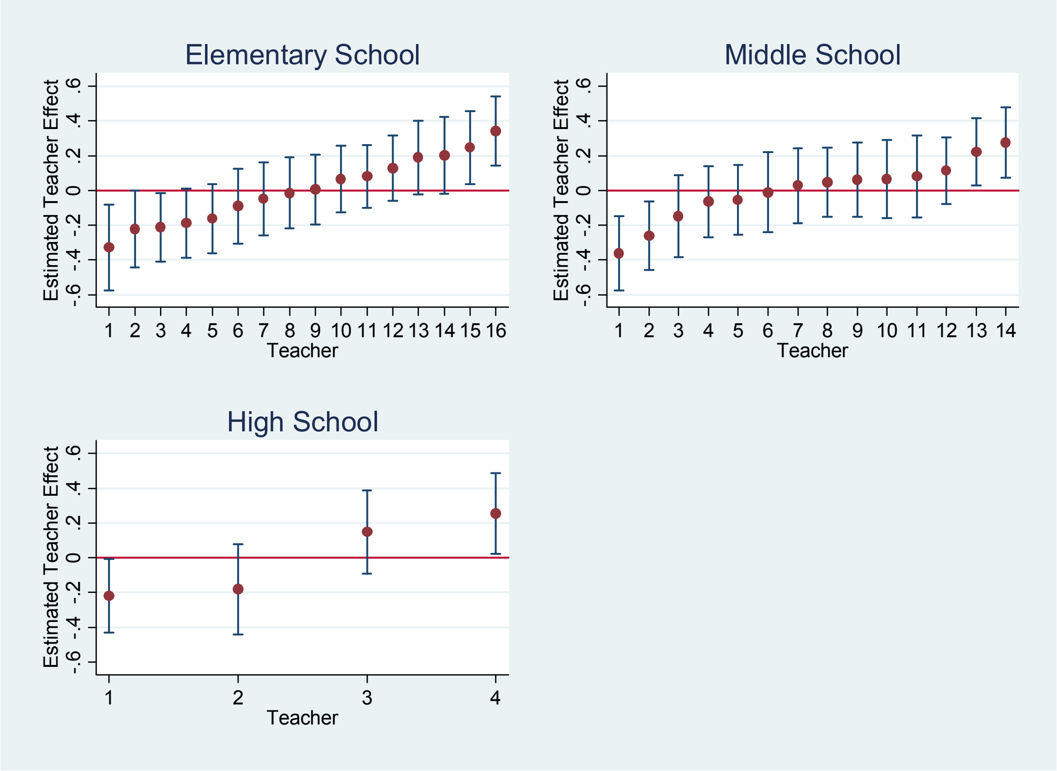
Source: Data collected and analyzed by Mathematica Policy Research.

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.

For the teacher model, the most meaningful comparisons are between teachers within the same school. Because the rankings measure the net influence of the classroom plus school characteristics (such as the effect of the principal or school culture on student achievement), it is not possible to disentangle the effect of the teacher from that of the school. To illustrate the comparison of teachers within a school, Figure 7 shows the teacher scores, along with the upper and lower bounds of a 90 percent confidence interval, for an award-winning elementary school, middle school, and high school.

As Figure 7 shows, the lower bound of the top-ranked teacher in each school is greater than zero but lower than the estimated score of the second-best teacher. For each grade level, the top-ranked teacher clearly is scoring higher than the lowest-ranked teacher, but teachers ranked near each other generally are not statistically distinguishable.

Figure 7. 90% Confidence Intervals for Two-Year Full VAM Teacher Estimates, for Sample Award-Winning Schools



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 teacher VAM estimate, which is given as the middle point.

REFERENCES

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

Booker, Kevin, Scott M. Gilpatric, Timothy Gronberg, and Dennis Jansen. "The Impact of Charter School Attendance on Student Performance." *Journal of Public Economics*, vol. 91, nos. 5-6, 2007, pp. 849-876.

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, 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.

Hoxby, Caroline. "Adequate Yearly Progress: Refining the Heart of the No Child Left Behind Act." In *Within Our Reach: How America Can Educate Every Child,* edited by John E. Chubb. New York: Rowman and Littlefield, 2005.

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.

Potamites, Elizabeth, Duncan Chaplin, Eric Isenberg, and Kevin Booker. "Measuring School Effectiveness in Memphis–Year 2." Washington, DC: Mathematica Policy Research, 2009.

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.

Sass, Tim R. "The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy." Washington, DC: The Urban Institute, Center for Analysis of Longitudinal Data in Research, 2008.

APPENDIX A: TECHNICAL DETAILS OF THE VALUE-ADDED MODEL

A. Estimation Sample

Most charter schools included in the final model provided Mathematica with test scores for their students from 2006–07 and 2007–08, as well as for at least one prior year. All schools that provided data for at least 15 students' current and past test scores were included in the model. Of the 83 elementary schools, the average number of students included in the model was 277 (the minimum at any school was 19 and the maximum was 1,133, including both the 2006–07 and 2007–08 school years). In the 37 middle schools, the average number of students in the model was 254 (with a minimum of 47 and a maximum of 1,016), including both years. In the 25 high schools, the average—including both years—was 148 students (with a minimum of 15 and a maximum of 1,398).[10]

For the teacher model, only teachers who could be linked to at least 15 students with non-missing current and past test scores were included. For elementary schools, 572 teachers at 63 schools were included, with a total of 25,283 student-test years in the two-year model. For middle schools, 233 teachers at 30 schools were included, with 12,686 student-test years. For high schools, 103 teachers at 21 schools were included, consisting of 4,270 student-test years. Not every school that was included in the school model had sufficient data linking students to teachers for inclusion in the teacher model. The teacher model includes students at award-winning schools and other schools that submitted sufficient usable data on student-teacher links.

B. Constructing School and Teacher Dosage Variables

For its model, Mathematica 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.[11] 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.

The teacher VAM incorporates enrollment data linking each student to each teacher. For each student-teacher pair, these data identify how many days each student was enrolled in each teacher's

[10] Looking at only the most recent year of data (the 2007–08 school year), elementary schools had an average of 157 students in the model (with a minimum of 15 and a maximum of 707), middle schools had an average of 145 (minimum 28 and maximum 617), and high schools had 110 students on average (minimum of 15 and a maximum of 676). Both these numbers and the numbers in the text represent the number of math scores that we were able to include for each school. The vast majority of schools had almost the same number of reading scores as math scores. One school in Missouri reported only math scores.

[11] Schools were requested to report the "number of days a child was considered a student at the school irrespective of whether or not the child was in attendance." The majority of schools (124 out of 145) reported that the maximum days enrolled for any student was between 165 days and 190 days, but 21 schools reported a maximum greater than 282 days and 2 of those reported a maximum of 365 days. To deal with this variation in maximums and also to attribute students who were at schools for the great majority of the year wholly to that school, we calculated each student's percentage of time at that school as days enrolled, divided by the 80th percentile of the distribution of days enrolled at that school. Fractions greater than one were top-coded to one. Another notable difference in the way that schools interpreted this question is that 57 schools reported the same number of days enrolled for all of their students and 6 schools did not report days enrolled for any of their students. In both cases, all students at these schools were assigned dosage values of one.

classroom, as well as what subject was taught. Using these enrollment data, Mathematica created dosage variables for each student, corresponding to the proportion of the year that student spent with that teacher. The dosage variable is set to zero for students who spent fewer than two weeks in a teacher's classroom and to one for students who spent all but two weeks or fewer in a given classroom. These teacher dosage values are then merged into the student-level data used in the school VAM.

For some students, the student-teacher link data fail to identify a teacher for at least one subject. Mathematica assigns students not linked to a teacher to a generic grade-school dosage variable for that school and grade combination. These grade-school dosage variables are included with the teacher dosage variables in the VAM estimation.

Teachers with very few students are not included in the teacher VAM. These primarily are teachers who did not teach at the school for the entire year or who taught students for only a small portion of the year. These teachers are identified by the sum of their dosage across all students. If a teacher had 10 or fewer dosage-weighted students, the dosage variable for this teacher is set to zero. The model implicitly estimates the effect of all omitted teachers as a single other teacher for that subject and grade.

C. Standardizing Test Scores

To compare student performance across different states, grades, subjects, and years, Mathematica standardizes the test scores by subtracting the statewide mean from each student's score and dividing each score by the statewide standard deviation, where means and standard deviations are calculated separately for each state, grade, subject, and year.[12] Mathematica 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.[13] The process first estimates each state's NAEP mean and standard deviation in each grade, subject, and year, using the fourth and eighth grade NAEP averages for that state in 2002–03, 2004–05, and 2006–07,[14] 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.

[12] Only New York was unable to provide statewide standard deviations by grade for 2007–08. To impute the standard deviations of New York state test scores, we used a prediction generated from a regression model. Using data from grades three to eight for reading tests in 2005–06 and 2006–07 and math in 2006–07 (n=18), we regressed the standard deviation of state test scores on the test score mean, the percentage of students scoring at each of four achievement levels, the cut-point score for the highest achievement level, an indicator variable for whether it was a math or reading test, and indicators for each grade level. The cut-point scores for the lower two achievement levels did not vary by grade, so they were not used because they were collinear with the grade fixed effects. The adjusted R2 was 0.84. Using the coefficients from this model and data available from 2007–08, we imputed the missing standard deviations for grades three to eight in reading and math.

[13] Hoxby (2005) 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.

[14] To estimate state NAEP means and standard deviations for untested grades and years, Mathematica 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.

The full one-year school model estimated without the NAEP adjustments produces results fairly similar to the model with the NAEP adjustments. The correlation of VAM estimates with and without the NAEP adjustment is 0.97.

D. Controlling For Measurement Error

One of the key control variables in the VAM is the student's prior year test score. 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 school model, using both subjects and the most recent year of data, the coefficient on the prior test score variable increases from 0.66 using ordinary least squares (OLS) to 0.84 using the IV. There is a similar increase for all other models. 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.[15]

E. The Value-Added Model

The VAM equation used to estimate school impacts is

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

where, $Y_{i,j,t}$ is the test score for student *i* in subject *j,* year *t*, $Y_{i,j,t-1}$ is the predicted value for the prior test score for student *i* in subject *j* and year *t-1*,[16] $X_{i,t}$ 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 $D_{i,t}$. All observations were weighted equally and the VAM is run jointly on all schools (elementary, middle, and high).

The VAM equation used to estimate teacher impacts is the same as the school model, except that in the teacher model, $D_{i,t}$ is a vector of teacher dosage variables. The vector $D_{i,t}$ includes one variable for each teacher in the model. Each variable equals the percentage of the year student *i* was in a classroom with that teacher. The value of any element of $D_{i,t}$ is zero if student *i* did not attend a classroom with that teacher. The teacher performance measures are the coefficients on $D_{i,t}$. All

[15] Davidson and MacKinnon (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.

[16] Unlike elementary and middle school students, high school students often have prior test scores that are more than one year prior. Also, previous to 2006, Massachusetts tested math in eighth grade and English language arts in seventh, so for those students, the prior math score is from a different year than the prior English language arts score. Mathematica 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.

observations were weighted equally. The VAM is run jointly for teachers in elementary and middle school grades, and then separately for high school teachers.[17]

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. We can include only variables for which data are available, so there may be other variables omitted that are still biasing our estimates. In addition to the student's lagged test score, the dosage variables, and a constant term, the VAM regressions include the following student-level variables:[18]

- Gender indicator[19]
- Race/ethnicity indicators (white, African American, Hispanic, Asian, Native American)
- Free or reduced price lunch indicator[20]
- Limited English proficiency (LEP) indicator[21]
- Special education status indicator
- First year at new school indicator
- Indicators for skipping or failing a grade since the last test

Other controls are grade level, subject, year indicators, and interaction terms, and a flag to indicate if we had to make an educated guess about which testing scale was used in Florida.[22]

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 or teacher performance measures are adjusted by student for the clustering of observations.

[17] The separation of high school teachers was due to computational issues occurring when including all the teachers in one joint model.

[18] 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.

[19] Gender also was interacted with subject.

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

[21] LEP also was interacted with subject.

[22] Some Florida charter schools submitted developmental scale scores for the FCAT; others submitted scores on the "regular" 100–500 point scale. Unfortunately, the scales overlap for reading in grades three to five and for math in grade three. For scores that fell in the range of both scales, we assigned a test type based on the type of test submitted for other students in that grade in the same school, for other grades in the same school, or for previous years in the same grade at the same school. In a very few cases, some students in a particular grade within a school clearly had one type of score, one or two students clearly had the other type of score, and some students had a score that fell within both ranges. In these cases, we assigned those with scores in between the ranges to the test type into which the majority of students in the grade at that school fell and created the flag mentioned in the text, indicating the need to make an educated guess about the scale.

Of the 145 schools in the analysis sample, 116 have data from both the 2006–07 and the 2007–08 school years that we were able to match with baseline achievement data from previous years. Twenty-nine schools have performance data from the most recent year only. We estimated value-added models based on performance data from 2007–08 only (the one-year model, which controls for baseline performance in a previous year) and also used performance year data from both 2007–08 and 2006–07 when available (the two-year model, which also controls for baseline performance data from previous years, resulting in up to three years of data being utilized). Data from 2007–08 most typically are matched with a baseline score in 2006–07. To avoid possible complications from 2006–07 observations appearing both as dependent and independent variables in the same model, we followed a two-step procedure recommended to us by Rob Meyer of the Wisconsin Center for Education Research to estimate the two-year models.

1. We estimated our VAM model, including instrumenting for the baseline test score, separately by year to get the coefficient on the baseline (B_1, the estimated value of β_1) and then calculate the

   "adjusted gain$_i$" = $Y_{i,j,t} – B_1 * Y_{i,j,t-1}$

2. Given the high level of precision at which B_1 typically is estimated (t=85.4 and 77.0 in years 1 and 2, respectively), we ignore the fact that the adjusted gains are estimated outcomes in the second step and regress the adjusted gain on all variables in the VAM model in the first step (except for the baseline test score), controlling for year and interacting year with subject and grade indicators (and the subject-grade interaction terms). The coefficients on the dosage variables from this step are the (un-shrunken) two-year VAM estimates.

Even using two years of performance year data, we still expect a good deal of noise in our estimates. To improve this situation, we used an empirical Bayesian method suggested by Morris (1983) and Carlin and Louis (1996). The basic idea is that imprecisely estimated school effects 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).[23] 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. This shrinkage estimator was used on all estimates, both one- and two-year models.

The mean weight placed on the overall mean from the full one-year VAM estimates is 0.15 (the minimum weight across schools is 0.10 and the maximum is 0.55).[24] The correlation between the

[23] Essentially, it is an iterative process that converges on the shrinkage weights for each school and the overall mean toward which all schools are shrunk jointly. 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, the overall mean and variance of the school effects are recalculated. We then calculate new weights using the new variance of the overall mean and repeat until the process converges.

[24] For the two-year model, the mean weight placed on the overall mean is 0.07 (the minimum weight across schools is 0.02 and the maximum is 0.55).

standard error of each school's estimate and the weight placed on the overall mean is 0.99. The correlation between the estimates before and after shrinking is 0.99.

We use the same shrinkage method for the teacher model but shrink each teacher's estimate toward the mean effect of all teachers at that school. Thus, each teacher's effect is the weighted average of the individual estimate and the school mean, where each teacher's weight depends on the standard error of the original estimate.

Both teacher and school models were estimated including a constant term. After shrinking the estimates, the coefficients were also mean-centered. An alternative model would omit the constant and include the omitted category as a control variable.[25] The difference between the value-added estimate of any individual school (teacher) relative to another is the same regardless of these modeling choices.

For each school, Mathematica 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 at that school. Using these standard errors, Mathematica calculates that the school effects are jointly significant. Mathematica 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 I.1 to I.3 show the confidence intervals for the rankings of schools in the elementary, middle, and high school categories for the full one-year model. Figures I.4 to I.6 show the same for the full two-year model.

An alternative way of describing the precision of the rankings is presented in Table A.1 below, which displays the ratios of the reliabilities of the estimates by grade level. Reliabilities measure 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. We also show the mean standard error divided by the standard deviation of the estimates, which can be interpreted as the percent of the variation due to noise.

Table A.1 shows that adding demographics to the model decreased the fraction of the variance due to noise and slightly increased the reliability for the elementary schools estimates, but had very negligible effects on the estimates from middle and high schools. The shrinkage estimator actually reduced the overall reliability of the results because, even though the precision (the standard errors) decreased, the variation in the estimates (σ) decreased more. Adding another year of data, when available, decreases the fraction of the variance due to noise and increases the reliability of estimates substantially, especially for middle and elementary schools.

An improvement in precision does not necessarily improve our ability to distinguish the performance of different schools from one another because the variation in performance also may decline. To test for this possibility, we compared the ratios of the average standard error to the standard deviations in performance measures. This ratio is .43 for the one-year model and .27 for the two-year model because the standard errors fell when adding a second year of data, while the

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

variation in the VAM estimates did not fall. Thus, adding an additional year of data does improve our ability to distinguish between the performance of different schools.[26]

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	Realibility (1-SSE/ σ^2)
Elementary Schools, N=83					
1 yr, no demographics, not shrunk	.085	.007	.192	.441	.802
1 yr, full, not shrunk	.084	.007	.197	.426	.815
1 yr, full, shrunk	.078	.006	.167	.466	.780
2 yr, full, shrunk	.047	.002	.164	.287	.911
Middle Schools, N=37					
1 yr, no demographics, not shrunk	.086	.008	.199	.429	.804
1 yr, full, not shrunk	.085	.008	.195	.435	.798
1 yr, full, shrunk	.078	.006	.160	.486	.755
2 yr, full, shrunk	.045	.002	.149	.306	.898
High Schools, N=25					
1 yr, no demographics, not shrunk	.108	.013	.369	.291	.905
1 yr, full, not shrunk	.107	.013	.362	.295	.904
1 yr, full, shrunk	.094	.009	.275	.341	.879
2 yr, full, shrunk	.072	.006	.305	.235	.939
All Schools, N=145					
1 yr, no demographics, not shrunk	.089	.008	.232	.383	.844
1 yr, full, not shrunk	.088	.008	.232	.380	.847
1 yr, full, shrunk	.081	.007	.187	.431	.808
2 yr, full, shrunk	.051	.003	.191	.267	.919

[26] Shrinkage also reduced our standard errors but did not improve our ability to identify effective schools because the standard deviation in performance dropped more than the standard error. The ratio of mean standard error to standard deviation in performance rises from 0.38 in the one-cohort model without shrinkage to 0.43 in the one-cohort model with shrinkage.

The precision of the teacher rankings is described in Table A.2 below, which displays the ratios of the reliabilities of the estimates by grade level. Table A.2 shows that, as with the school estimates, adding another year of data, when available, decreases the fraction of the variance in the teacher estimates due to noise and increases the reliability of those estimates.

Table A.2 Reliabilities of the Teacher 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 Teachers, N=572					
1 yr, full, shrunk	.159	.026	.224	.709	.291
2 yr, full, shrunk	.123	.021	.218	.563	.437
Middle School Teachers, N=233					
1 yr, full, shrunk	.146	.024	.217	.672	.328
2 yr, full, shrunk	.114	.019	.208	.546	.454
High School Teachers, N=103					
1 yr, full, shrunk	.183	.037	.290	.630	.370
2 yr, full, shrunk	.177	.036	.287	.617	.383
All Teachers, N=908					
1 yr, full, shrunk	.160	.034	.261	.613	.387
2 yr, full, shrunk	.132	.022	.235	.561	.439

MATHEMATICA
Policy Research, Inc.

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC