American Economic Review: Papers & Proceedings 2014, 104(5): 406–411
http://dx.doi.org/10.1257/aer.104.5.406

Using School Choice Lotteries to Test Measures of School Effectiveness[†]

By David J. Deming*

The measurement of school effectiveness is a central feature of educational accountability policies in all 50 US states and around the world. While school accountability measures are often based on test score levels (e.g., percent proficient), critics argue that test score gains are a fairer way to judge schools' contributions to student achievement (e.g., Ladd and Walsh 2002, Ryan 2004). Such "value-added" measures (VAMs) have now been introduced into the accountability regimes of at least 30 US states (Blank 2010). The growing interest in VAMs has given rise to a large literature which deals with technical issues such as model specification, choice of sample and outcome, and measurement error in the estimation of "school effects" (Raudenbush and Willms 1995; Meyer 1997; McCaffrey et al. 2003; Rubin, Stuart, and Zanutto 2004; Reardon and Raudenbush 2009). Yet random variation in school attendance is both rare and necessary to test the validity of VAMs, and to guide the selection of models for measuring causal effects of schools.

In this paper I use data from a public school choice lottery in Charlotte-Mecklenburg (CMS) to test the validity of school value-added models. Students were guaranteed assignment to their neighborhood school but could apply to attend other schools in CMS, with admission to oversubscribed schools determined by lottery. This yields random assignment to schools, albeit within a self-selected sample of applicants. I estimate a variety of school VAMs, varying the model specification, outcome, and sample on which VAM is calculated. I then use these nonexperimental estimates of "school effects" to predict the impact of winning the lottery to attend a chosen school on student achievement.

Overall, I find that VAMs are a remarkably accurate out-of-sample predictor of student achievement. In specifications with minimal controls (i.e., one year of prior test scores and no other covariates) and two or more years of prior data, I fail to reject the hypothesis that school effects are unbiased. VAMs with a richer set of covariates perform similarly.

A few existing studies of "teacher effects" find that conditioning on prior test scores and other characteristics is sufficient to account for sorting of students across teachers within a school (Kane and Staiger 2008; Chetty, Friedman, and Rockoff 2013; Kane et al. 2013). However, the assumptions of school VAMs require that the same covariates are sufficient to account for sorting of students *across schools*, which may be less likely to hold. A small existing literature compares the results from lottery-based admission to charter schools to results that use observational designs (Hoxby and Rockoff 2005; Abdulkadiroğlu et al. 2011; Deutsch 2012; Angrist, Pathak, and Walters 2013). This work connects to the broader tradition in economics of comparing experimental to nonexperimental evaluation methods, beginning with LaLonde (1986). As pointed out by Rothstein (2010), quantifying the bias that arises from commonly used VAMs has important implications for education policy. If high-stakes school accountability ratings are biased or inaccurate, they are unlikely to improve performance and may lead to wasteful compliance behavior (e.g., Baker 2002).

I. Data and Value-Added Models

The main data source for this article is a panel of administrative data on all students enrolled in CMS from 1996 to 2004. These data contain detailed information on student demographics,

*Harvard Graduate School of Education, Gutman 411, Appian Way, Cambridge, MA 02139 (e-mail: david_deming@gse.harvard.edu).

[†]Go to http://dx.doi.org/10.1257/aer.104.5.406 to visit the article page for additional materials and author disclosure statement(s).

enrollment histories by school, grade, and year, and end-of-year (EOY) test scores in math and English language arts (ELA). Students are tested in grades 3 through 8 every year and in both subjects. I use test scores from the 1996–1997 through 2001–2002 school years as the main covariates in our VAM estimation, and I use the 2002–2003 and later test scores as outcomes.

I estimate school value-added models (VAMs) of the general form

$$(1)\quad A_{ijt} = X_{ijt}\beta + \omega_{ijt};\quad \omega_{ijt} = \mu_j + \theta_{jt} + \varepsilon_{ijt}.$$

The dependent variable A_{ijt} is the state-standardized EOY score in math or reading for student i in school j and year t. X_{ijt} is a vector of student-level covariates, and ω_{ijt} is a residual term. The key parameter of interest is the "school effect" μ_j, which can be obtained either by computing the average school-level residual (i.e., random effects) or by direct estimation (i.e., fixed effects). VAMs rely on the covariate vector X_{ijt} to adjust for observed differences in student characteristics across schools. Specifically, if assignment to schools is uncorrelated with unobserved determinants of achievement conditional on the covariates in X_{ijt}, μ_j can be interpreted as the causal impact of attending school j relative to the average school for a randomly chosen student (e.g., Reardon and Raudenbush 2009). The addition of a lagged test score A_{ijt-1} into the X vector is what gives the model a "value-added" interpretation, since we are asking whether test score *gains* are higher in some schools than others. This setup closely follows the teacher effects literature (e.g., McCaffrey et al. 2004; Kane and Staiger 2008; Kane et al. 2013).

I introduce three sets of covariates into the X_{ijt} vector. The first includes a year fixed effect but no other covariates, and essentially ranks schools by average test scores (in levels). The second specification includes only a third-order polynomial in the prior year's reading and math scores. This model resembles the growth-based approach to school ratings used by several states. The third specification adds gender, race, free or reduced price lunch eligibility (a proxy for income based on the Federal poverty standard), and prior peer achievement. This emulates the traditional VAM approach used in previous work (e.g., McCaffrey et al. 2004; Kane and Staiger 2008).

In all models, I use data from prior years to predict the impact of attending a particular school in the year that the lottery was conducted (2002–2003). To empirically assess the importance of using multiple years of data, I estimate VAMs with only one prior year of data (2001–2002), two prior years of data (2000 to 2002), and then all five years that are available in the CMS panel (1997 to 2002). Three groups of covariates times three different samples equals nine different specifications with average test scores in the spring of 2003 as the main outcome.

Like Kane et al. (2013), I find that average residual and fixed effects approaches produce very similar results, and so I report only the results using the average residual (i.e., random effects) approach.[1] I estimate equation (1) separately for grades 4 through 8, effectively obtaining "school-by-grade" effects using multiple years of data.[2] The main outcome of interest is the average of a student's standardized math and ELA score at the end of the indicated school year.[3]

Prior work on teacher effects has employed Empirical Bayes (also called shrinkage) estimators, which attenuate teacher effects toward zero based on the amount of year-to-year variation in the estimate (e.g., McCaffrey et al. 2004; Kane and Staiger 2008; Kane et al. 2013).[4]

[1] Fixed effects models account for correlation between μ_j and the covariates in X_{ijt}. However, since most of the variation in achievement is within schools rather than between schools, models with fixed effects and average residuals produced school effects that are correlated about 0.95 with a full set of covariates, and greater than 0.99 in more basic specifications.

[2] I also pursue an alternative approach that estimates a single school effect across multiple grades. Those results, which are available upon request, are similar in magnitude to the main results but much noisier.

[3] While I can also estimate separate VAMs for math and ELA, the average score is preferable for two reasons. First, it increases precision. Second, and most important, school effectiveness is very likely to "spill over" across tests. As evidence, I note that the cross-subject, within-year correlation between math and ELA school effects is usually between 0.5 and 0.6. This is almost always larger than the within-subject, across-year correlation, which averages closer to 0.4 for math and 0.25 for reading. Separate results for math and reading are available upon request.

[4] We follow the procedure used in Kane and Staiger (2008), and in Chetty, Friedman, and Rockoff (2013) for the special case where μ_j is assumed to be fixed over time and θ_{jt} and ε_{ijt} are i.i.d. In this case μ_j is multiplied by the signal-to-noise ratio or "reliability," which is essentially the

Chetty, Friedman, and Rockoff (2013) modify this approach by allowing for a nonparametric autocovariance structure over past years of data, essentially allowing teacher effects to "drift" over time.[5] I report unshrunken school effects as well as results from both methods of adjustment.

II. Comparison of Lottery Results to VAM Estimates

Here I provide only a very brief description of the CMS school choice lottery—for more details, see Hastings, Kane, and Staiger (2010); Deming (2011); or Deming et al. (2014). Parents submitted their top three choices for other schools and were guaranteed admission to the neighborhood school. Admission was determined by random numbers within each lottery, defined at the school-grade-priority group level, with a small number of different priority groups based on factors like sibling attendance. My analysis focuses on a sample of 2,599 students in 118 separate lotteries with "marginal" priority groups. For these students, (i) the probability of admission was neither zero nor one, and (ii) assignment was determined only by a random number. Online Appendix Table A1 compares the lottery sample to other students in CMS. Lottery applicants are fairly representative of their classmates on observed characteristics, although they are somewhat more likely to be African-American and have modestly lower test scores.

I compare the actual impact of winning the lottery to the predicted impact that is implied by the school VAMs estimated in Section I above. To do this, I estimate

$$(2)\quad VAM_{ij}^{A} = \theta W_{ij} VAM_{ij}^{1} + \gamma(1 - W_{ij}) VAM_{ij}^{N} + \pi X_{ij} + \Gamma_j + \varepsilon_{ij}$$

$$(3)\quad A_{ij} = \delta \widehat{VAM_{ij}^{A}} + \beta X_{ij} + \Gamma_j + \varepsilon_{ij}.$$

Where VAM_{ij}^{A} is the VAM estimate for school j attended by student i in the fall of 2002, VAM_{ij}^{1} is the VAM estimate for the student's first choice school, VAM_{ij}^{N} is the VAM estimate for the student's neighborhood or "home" school, and W_{ij} is an indicator variable that is equal to one if student i has a winning lottery number for admission to school j. A_{ij} are end-of-year (EOY) test scores in spring 2003, X_{ij} is a vector of prelottery covariates that is included only for improved precision, Γ_j is a set of lottery fixed effects, and ε_{ij} is a stochastic error term.[6] To see the intuition for this specification, imagine that a student applies to a school which has an estimated "value added" that is 0.1 standard deviations (SDs) higher than their outside option, usually the neighborhood school.[7] If the VAM estimate is a causal measure of the school's impact on achievement, a student who wins the lottery will score 0.1 SDs higher on the test at the end of the year. In that case, the δ coefficient in equation (4) will have a value of exactly one, because the actual estimate exactly matches the impact on achievement that is predicted by the VAM estimate. Likewise, if the actual impact on achievement is somewhat less (say 0.05 SDs), the coefficient may be significantly greater than zero but also significantly less than one, implying some upward bias in the VAM estimate.

There are at least three reasons why VAM estimates may not predict the impacts of winning the lottery. First, VAMs may be biased due to sorting on unobserved determinants of achievement (e.g., Rothstein 2010). Second, if "true" school effects vary over time independent of estimation error, then out-of-sample forecasts based on prior cohorts may be a poor predictor of future effectiveness. Third, since students in the lottery sample are self-selected, the impact

year-to-year variance in the school effect estimate divided by the total variance after correcting for school size.

[5] In Kane and Staiger (2008) and other studies that use the Empirical Bayes approach, all prior years of data are weighed equally. In Chetty, Friedman, and Rockoff (2013), the "shrinkage" factor is estimated using the autocovariance of mean test score residuals across years. This allows for some prior years of data to be weighted more heavily. They find that more recent years are more predictive of future teacher effects.

[6] Because the lotteries were conducted at the school-grade-priority group level, the number of lotteries is greater than the number of schools. I suppress subscripts for grade and priority group for notational convenience. The X_{ij} vector includes controls for race, gender, free or reduced price lunch, and a third-order polynomial in prior year (2001–2002) math and reading test scores plus indicator variables for missing scores.

[7] Most students who lost the lottery for their first choice ended up in their neighborhood school. However, we cannot observe the counterfactual school that lottery winners would have attended. As a sensitivity check, we construct counterfactual "control" schools using students' submitted choices combined with the ex post probability of admission. This procedure improves the accuracy of VAMs slightly.

of attending a school may be different for them than for a randomly chosen student from a prior cohort. Each of these reasons could lead to bias in either direction.

The main results of the paper are in Table 1. The first nine columns report results from different VAM specifications, unadjusted for "shrinkage"—three groups of covariates in the X_{ij} vector and three different estimations samples, as described in Section I. Columns 10 through 12 show results that employ both the Empirical Bayes shrinkage method used in past work such as Kane and Staiger (2008) and the autocovariance-adjusted "drift" procedure employed by Chetty, Friedman, and Rockoff (2013). I also report the standard deviation of the school effects estimates for each model. Because the VAM estimates are generated regressors, I block bootstrap the standard errors at the lottery level.

Columns 1 through 3 consider the accuracy of a VAM with no covariates at all, essentially asking whether winning the lottery to attend a school with higher test scores in levels increases student achievement. The coefficients are small and not significantly different from zero, suggesting that average test scores alone contain almost no information about a school's causal impact on achievement. Columns 4 through 6 show results from the "gains" model, which includes only one year of prior test scores in the X_{ij} vector. The performance of VAMs improves dramatically in these specifications. When we use only one year of prior data to estimate value added, controlling for prior scores increases the coefficient from 0.025 in column 1 to 0.531 in column 4. With two or more years of prior data, the VAM estimates are highly accurate (0.807 and 0.966 in columns 5 and 6, respectively), and we fail to reject the hypothesis that they are biased (i.e., statistically different from one). Adding demographic covariates to the VAMs leads to slightly larger coefficients than in the gains specification.

Columns 10 through 12 show the impact of "shrinkage" adjustments on the forecasting accuracy of VAMs. When VAM estimates are based on only one year of prior data, adjustment for shrinkage improves the accuracy of the forecast by about 18 percent (from 0.531 to 0.627). However, when I average the VAM estimates across multiple years, "shrinkage" adjustment actually reduces forecasting accuracy (1.237 in column 9 compared to 1.602 in column 11). The "drift" adjustment, as in Chetty, Friedman, and Rockoff (2013), produces results that are slightly more accurate than unshrunken estimates, and substantially more accurate than Empirical Bayes shrinkage.[8]

Online Appendix Table A2 tests for the persistence of school effects estimates by comparing the VAM for a student's assigned school and grade to achievement outcomes in spring 2004, two years after the lottery. The second year VAM estimate is highly predictive of second year achievement. Moreover, when I include both the first and second year estimates together, I find evidence that first year school effects have a persistent impact on second year achievement. However, the coefficients are imprecisely estimated.

Overall, I find that VAM estimates line up very closely with estimates from lottery-based random assignment. For most commonly used VAM specifications, I cannot reject the hypothesis that school effects are unbiased predictors of actual achievement. While this study offers hope that value-added modeling can be used to make inferences about school effectiveness, I conclude with a few cautionary notes.

First, while VAMs appear to be an unbiased predictor of student achievement, many other important outcomes of schooling are not measured here. Schools and teachers that are good at increasing student achievement may or may not be effective along other important dimensions (e.g., Chetty, Friedman, and Rockoff 2013; Deming et al. 2013; Jackson 2012). Second, this study uses a relatively small sample from a single school district. The proliferation of public school choice and charter school lotteries across the United States provides an opportunity for researchers to test the accuracy of VAMs in

[8] Both methods of shrinkage adjustment attempt to estimate a time-invariant school effect μ_j. Yet if some share of the yearly variance θ_{jt} in school effects comes from true differences in effectiveness, then shrinkage that is based on the autocovariance of estimates will make school effects estimates too small. In the teacher effects literature, Chetty, Friedman, and Rockoff (2013) refer to this source of variation as "teacher bias." When I use all prior years of data to form the VAM estimate, the empirical bayes procedure used in Table 1 yields a reliability estimate of about 77 percent (1.237/1.601). The results here imply that a reliability of 93 percent produces the best forecast (i.e., the coefficient that is closest to 1). However, this particular reliability estimate may not hold in other samples.

Table 1—Validating Models of "School Effects" Using Lottery Data

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Outcome is spring 2003 test scores						
School "value-added" in 2003	0.025 [0.077]	0.034 [0.070]	0.014 [0.072]	0.531* [0.208]	0.807** [0.236]	0.966** [0.342]
Years of prior data	2002 only	2001–2002	1998–2002	2002 only	2001–2002	1998–2002
Covariates in VAM	None	None	None	prior scores (1 lag)	prior scores (1 lag)	prior scores (1 lag)
Shrinkage adjustment	None	None	None	None	None	None
p-value on $F(VA = 1)$	0.000	0.000	0.000	0.024	0.413	0.920
SD of school effects	0.431	0.417	0.397	0.110	0.096	0.073
Observations	2,599	2,599	2,599	2,599	2,599	2,599
	(7)	(8)	(9)	(10)	(11)	(12)
Panel B. Outcome is spring 2003 test scores						
School "value-added" in 2003	0.547** [0.194]	0.908** [0.231]	1.237** [0.347]	0.627** [0.207]	1.602** [0.450]	1.185** [0.323]
Years of prior data	2002 only	2001–2002	1998–2002	2002 only	1998–2002	1998–2002
Covariates in VAM	demogs + prior scores	demogs + prior scores	demogs + prior scores	demogs + prior scores	demogs + prior scores	demogs + prior scores
Shrinkage adjustment	None	None	None	Empirical Bayes	Empirical Bayes	"Drift" adjustment
p-value on $F(VA = 1)$	0.020	0.689	0.495	0.071	0.180	0.567
SD of school effects	0.102	0.088	0.067	0.088	0.047	0.054
Observations	2,599	2,599	2,599	2,599	2,599	2,599

Notes: Each column shows results from an estimate of the two-stage least squares (2SLS) system in equations (2) and (3) in the paper, where the "value-added" model (VAM) estimate in a student's fall 2002 school is the first-stage endogenous variable, and the instrument is the VAM estimate in the first choice school for lottery winners and the VAM estimate in the neighborhood school for lottery losers. The outcome in each regression is the average of students' spring 2003 math and reading scores. The reported coefficients will, thus, be equal to one if the VAM indicated in each column is a perfect predictor of the impact of attending a student's first choice school on student achievement. Columns 1 through 9 report results from three different choices of prior covariates and three different estimation samples—see the indicated column and the text for details. Columns 10 through 12 report results from "shrinkage"-adjusted VAMs—see the text for details. For each regression, I report the p-value on an F-test of the hypothesis that the coefficient is "unbiased," i.e., equal to one. Standard errors are block bootstrapped at the lottery level.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

other settings. Finally, it should be noted that the "effects" of schools on student achievement arise from a combination of factors, only some of which are under the school's control. VAM estimation does not uncover the mechanisms that underlie the production of student achievement, and variables such as peer influence and school context may have important influences independent of the school's actions (Raudenbush and Willms 1995; Todd and Wolpin 2003). For all these reasons, we should be cautious before moving toward policies that hold schools accountable for improving their "value added."

REFERENCES

Abdulkadiroğlu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, and Parag Pathak. 2011. "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots." *Quarterly Journal of Economics* 126 (2): 699–748.

Angrist, Joshua D., Parag Pathak, and Christopher R. Walters. 2013. "Explaining Charter School Effectiveness." *American Economic Journal: Applied Economics* 5 (4): 1–27.

Baker. George P. 2002. "Distortion and Risk in Optimal Incentive Contracts." *Journal of Human Resources* 37 (4): 728–51.

Blank, Rolf K. 2010. "State Growth Models for School Accountability: Progress on Developing and Reporting Measures of Student Growth." Council of Chief State School Supervisors: Washington, D.C.

Chetty, Raj, John N. Friedman, and Jonah Rockoff. 2013. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value–Added Estimates." National Bureau of Economic Research Working Paper 19423.

Deming, David J. 2011. "Better Schools, Less Crime?" *Quarterly Journal of Economics* 126 (4): 2063–2115.

Deming, David J., Justine S. Hastings, Thomas J. Kane, and Douglas O. Staiger. 2014. "School Choice, School Quality, and Postsecondary Attainment." *American Economic Review* 104 (3): 991–1013.

Deming, David J., Sarah R. Cohodes, Jennifer Jennings, and Christopher Jencks. 2013. "School Accountability, Postsecondary Attainment and Earnings." National Bureau of Economic Research Working Paper 19444.

Deutsch, Jonah. 2012. "Using School Lotteries to Evaluate the Value-Added Model." Unpublished.

Hastings, Justine S., Thomas J. Kane, and Douglas O. Staiger. 2010. "Heterogeneous Preferences and the Efficacy of Public School Choice." Unpublished.

Hoxby, Caroline M., and Jonah E. Rockoff. 2005. "The Impact of Charter Schools on Student Achievement." Unpublished.

Jackson, C. Kirabo. 2012. "Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina." National Bureau of Economic Research Working Paper 18624.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment.* Seattle: Bill & Melinda Gates Foundation.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." National Bureau of Economic Research Working Paper 14607.

Ladd, Helen F., and Randall P. Walsh. 2002. "Implementing Value-Added Measures of School Effectiveness: Getting the incentives right." *Economics of Education Review* 21 (1): 1–17.

Lalonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review,* 126 (4): 604–20.

McCaffrey, Daniel F., Lockwood, J. R., Koretz, Daniel, Louis, Thomas A., and Laura Hamilton. 2004. "Models for Value–Added Modeling of Teacher Effects." *Journal of Educational and Behavioral Statistics* 29 (1): 67–101.

Meyer, Robert H. 1997. "Value-added Indicators of School Performance: A Primer." *Economics of Education Review* 16 (3): 283.

Reardon, Sean F. and Stephen W. Raudenbush. 2009. "Assumptions of Value–Added Models for Estimating School Effects." *Education Finance and Policy* 4 (4): 492–519.

Raudenbush, Stephen W. and J. Douglas Willms. 1995. "The Estimation of School Effects." *Journal of Educational and Behavioral Statistics* 20 (4): 307–35.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125 (1): 175–214.

Rubin, Donald B., Elizabeth A. Stuart; Elaine L. Zanutto. 2004. "A Potential Outcomes View of Value–Added Assessment in Education." *Journal of Educational and Behavioral Statistics* 29 (1): 103–16.

Ryan, James E. 2004. "The Perverse Incentives of the No Child Left Behind Act." *NYU Law Review* 79: 932–89.

Todd, Petra E. and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal* 113 (485): F3–F33.