Abstract Title Page
Not included in page count.

Title:

Using School Lotteries to Evaluate the Value-Added Model

Authors and Affiliations:

Jonah Deutsch
University of Chicago

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

There has been an active debate in the literature over the validity of value-added models. Rothstein (2010) finds that the value-added estimate of a student's teacher in year *t* predicts that student's test score gain in year *t-1*. He concludes that this evidence of selection based on past gains violates the assumptions of value-added models, leading to potentially significant bias. Kane and Staiger (2008) employ an experimental evaluation to test for systematic bias. They estimate teacher value-added using several years of observational data, and then use these estimates to predict student performance when students are randomly assigned to classrooms within pairs of teachers. They find that the bias is insignificant, although they cannot reject substantial bias due to their small sample size and access to only a select group of teachers.

Chetty et al. (2011) first test for the importance of selection on unobservables by introducing a set of variables on parental characteristics absent from most data. After finding that these variables are not correlated with value-added estimates, they test for bias using teachers who switch schools and conclude that they can rule out plausible types of selection that would bias estimates of teacher quality.

Hoxby and Rockoff (2004) also compare lottery-based estimates of charter school effects to those from a value-added model. However, as they point out, the students used in the value-added approach is a select sample, limited to those who switch schools in late elementary grades. Furthermore, their value-added model is a difference-in-difference model, and is thus quite different from the models used more commonly in the value-added literature. Deming et al. (2011) compare lottery-based results from a school choice initiative to school value-added estimates, though their test is of a joint null that the value-added model is correctly specified and other assumptions about how students and parents behave in a school choice system.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

I test the central assumption of value-added models that school assignment is random relative to expected test scores conditional on prior test scores, demographic variables, and other controls. I use a charter school lottery to identify school effects, and then compare this "experimental" estimate to that of a school value-added model, which is estimated from all students in a large district. I use an innovative approach to generate a value-added-based estimator of the lottery effect that is unbiased under the null hypothesis that the value-added model is correctly specified. My results imply small but potentially important upward bias in the math, and very minimal bias in reading.

Setting:

Description of the research location.
(May not be applicable for Methods submissions)

The data used in this analysis come from a charter school in Chicago. As mandated by Illinois law, the charter school holds a lottery to determine admission when the number of

applicants exceeds the number of open slots. The students in these lotteries are then matched to administrative records from the Chicago Public Schools (CPS).

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

One novelty of this paper is the formal linkage of the lottery-based and value-added model, such that under the assumptions of the value-added model (along with a standard exclusion restriction), the Intent-to-Treat (ITT) effect can be shown to be a combination of parameters of the value-added model. This allows for a cleaner comparison, whereby we can be precise about the assumptions necessary to make the two estimates equal in expectation. I can also then conduct a formal test of their equality.

I also investigate the trade-off in combining many years (and grades) into the model to estimate school effects. In the teacher value-added literature, McCaffrey et al. (2009) has argued that combining several cohorts of students to estimate teacher effects mitigates sampling error. However, there may be a trade-off if true teacher performance varies significantly from year to year and we wish to capture that performance. That is, even if we were able to completely eliminate sampling error, there would still presumably be year-to-year differences in effects within teachers, due to varying levels of performance. It's not clear that a policy-oriented metric would want to average out those differences. As Hanushek and Rivkin (2010) point out, combining multiple years may diminish the role value-added accountability systems can play in incentivizing productive behaviors. As an example, imagine an accountability system with some form of incentive (either positive or negative) for performance in year *t*. In a world without sampling error, a teacher's effort in year *t* will have less of an influence on their evaluation if the metric averages over years *t, t-1, t-2, . . . , etc.* than if it is based on year *t* alone.

In general, the literature has not paid close attention to this distinction, though it is clearly important in using these models for accountability purposes. Indeed, Deming et al. (2011) focus on school effects for one year and grade in their test score analysis, and indicate this is because of large changes in schools from one year to the next driven by a district policy change. But district-level reforms, school-level policy and personnel changes occur quite frequently, and averaging over years and/or grade levels may be hiding differences in performance that are relevant to policy-makers.

A similar argument can be made over whether to aggregate over several grades when identifying school effects. With different personnel working at different grade levels, combined with different curriculums and pedagogy, information about 4th graders may tell us little about how effectively a school teaches 8th graders.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

Consider a standard, fixed effects value-added model:

$$(1) \quad Y_{ist} = Y_{it-1}\gamma + X_i\beta + \sum_{s=1}^{K} I_{it}^{s}\,\pi_s + \varepsilon_{ist}$$

Where Y_{ist} is the test score of student *i* who attended school *s* at time *t*, $Y_{it-1}$is the lagged test score, X_i includes time-invariant student characteristics, I_{it}^{s} is an indicator for student *i* attending school *s* at time *t*, π_s is the effect of attending school *s*, and ε_{ist} is the white noise error term.

Now consider the typical model identifying the Intent-to-Treat (ITT) effect of winning a lottery:

$$(2) \qquad Y_{ist} = \mu + A_{it}\delta + u_{ist}$$

Where A_{it} is an indicator equal to 1 if student *i* at time *t* won the lottery and was admitted to the charter school, and 0 otherwise.[1] Due to the independence between A_{it} and u_{ist}, we can then write:

$$\delta = E[Y_{ist}|A_{it} = 1] - E[Y_{ist}|A_{it} = 0]$$

This will lead us to an equation for δ, the ITT effect, in terms of the parameters of our value-added model when we replace the Y_{ist} above with the outcome from (1):

$$(3) \qquad \delta = \sum_{s=1}^{K} \pi_s E[I_{it}^s|A_{it} = 1] - \sum_{s=1}^{K} \pi_s E[I_{it}^s|A_{it} = 0] + E[\varepsilon_{ist}|A_{it} = 1] - E[\varepsilon_{ist}|A_{it} = 0]$$

Note that the first two terms of (1) drop out, as they are determined prior to random assignment and are independent of A_{it}. We can then invoke the familiar exclusion restriction used in the instrumental variables literature that winning or losing the lottery affects an individual's outcome only through the lottery's effect on determining the school the individual attends: $\varepsilon_{ist} \perp A_{it}$. This assumption makes the final two terms in (3) drop out: [2]

$$(4) \qquad \delta = \sum_{s=1}^{K} \pi_s E[I_{it}^s|A_{it} = 1] - \sum_{s=1}^{K} \pi_s E[I_{it}^s|A_{it} = 0]$$

$$= \sum_{s=1}^{K} \pi_s (p_{1s} - p_{0s})$$

where $p_{js} = E[I_{it}^s|A_{it} = j], j = 0,1$. In other words, p_{js} is the conditional probability of attending school *s*. The experimental causal estimate now appears as a function of parameters from the value-added model.

To generate a value-added based estimate of the lottery effect, I replace π_s with $\hat{\pi}_s$ from the value-added model, and p_{js} with another estimate. In the case of the latter, notice that $p_{1s} - p_{0s}$ is equal to α_1^s from the following model:

$$(5) \qquad I_{it}^s = \alpha_0^s + A_{it}\alpha_1^s + e_{it}$$

Further, the random generation of A_{it} ensures that the OLS coefficient $\hat{\alpha}_1^s$ is an unbiased estimate of α_1^s. Thus we can replace $p_{1s} - p_{0s}$ with $\hat{\alpha}_1^s$, arriving at the following estimator:

$$(6) \qquad \hat{\delta}_{VA} = \sum_{s=1}^{K} \hat{\pi}_s \hat{\alpha}_1^s$$

Arithmetically, (6) is exactly equivalent to the estimate one would get by estimating (2) with the value-added estimate of the school attended by each student replacing their post-lottery test score.

Usefulness / Applicability of Method:

Demonstration of the usefulness of the proposed methods using hypothetical or real data.

See tables below.

[1] Again, for ease of exposition, consider the case where a student enters a maximum of 1 charter school lottery.

[2] Note that we do not require the second central IV assumption: monotonicity.

Findings / Results:

Description of the main findings with specific details.
(May not be applicable for Methods submissions)

None of the biases are statistically significant, though that may be too low a standard. For example, a naïve value-added-based hypothesis that all schools attended by winners and losers were the same would also not be rejected by these tests (using the same standard errors of the differences estimated in the table). Many of the estimates are almost as close to this naïve hypothesis as they are to the true lottery estimate. Examining them this way, the biases seem important from a policy perspective. Another way to put the bias in context is in relation to the standard deviation of school effects.

The standard deviations of school effects are roughly consistent with the literature, with the possible exception of the random effects model. The fact that the estimated "true" variance of school effects for the first two pooled fixed effects models is negative may call into question the utility of those metrics. Since any true variance cannot be negative, it seems inappropriate to employ the estimated variance in the empirical Bayes framework, so I present the raw, de-meaned school effects estimates instead. Nevertheless, it may be that these models do not adequately identify school effects if the unadjusted variance of the school effects is less than the mean of the squared standard errors.

For math, all but one of the value-added estimates is greater than the lottery-based estimate, lending support to the hypothesis that value-added fails to control for unobserved components of achievement that are positively correlated with selection into a charter school. The fact that all of value-added models produce very similar estimates, save for the random effects model, implies that there is something systematic in the bias that is not related to any of the different estimation techniques. The random effects estimate reflects the least amount of bias of the estimates, and is very close to the lottery effect. The biases range from 0.027 to 0.77 standard deviations of school effects.

Within the average residual models, the degree of bias increases with the degree of pooling. That is, pooling over years and grades generates more bias than pooling over years, within grades, which generates more bias than estimating each year-by-grade combination separately.

The reading results (table 4) indicate that the value-added models are close to unbiased in reading. With the exception of the random effects model, which is biased downward, the differences between the value-added models and the lottery estimate are insignificant by any reasonable standard. The same caveat described above regarding the "true" variances applies to all three pooled fixed effects models in reading: these estimates are not generated with shrinkage. The school standard deviations tend to be smaller for reading than for math within each model, which is consistent with much of the literature (see Hanushek and Rivkin, 2010 for an overview).

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

I cannot reject the null that the value-added models are unbiased. In math, most of the models diverge from the experimental estimate, by as much as 77% of the standard deviation in school effects. Models that aggregate over years and grades perform worse than those that estimate each year-by-grade separately. In reading, the value-added models perform well, with very minimal divergence. In general, for both subjects, the different models align closely with one another.

Appendices

Not included in page count.

Appendix A. References

Aaronson, Daniel, Lisa Barrow, and William Sander (2007). "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25(1): 95-135.

Ballou, Dale, William Sanders, and Paul Wright (2004). "Controlling for Student Background in Value-Added Assessment of Teachers." *Journal of Educational and Behavioral Statistics* 29(1): 37-65.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2011). "The Long-term Impacts of Teachers: Teacher Value-added and Student Outcomes in Adulthood." NBER Working Paper 17699.

Deming, David J., Justine S. Hastings, Thomas J. Kane, and Douglas O. Staiger (2011). "School Choice, School Quality and Postsecondary Attainment." NBER Working Paper 17438.

Guarino, Cassandra M., Mark D. Reckase, and Jeffrey M. Wooldridge (2012). "Can Value-Added Measures of Teacher Performance be Trusted." Education Policy Center working paper 18.

Hanushek, Eric A. and Steven G. Rivkin (2010). "Generalizations about Using Value-Added Measures of Teacher Quality." *American Economic Review* 100(2): 267-71.

Hoxby, Caroline M. and Jonah E. Rockoff (2004). "The Impact of Charter Schools on Student Achievement." HIER Working Paper.

Jackson, C. Kirabo (2011). "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers." NBER Working Paper 15990.

Jacob, Brian A. and Lars Lefgren (2008). "Can Principals Identify Effect Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics* 26(1): 101-36.

Kane, Thomas J. and Douglas O. Staiger (2008). "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper 14607.

LaLonde, Robert J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4): 604-20.

McCaffrey, Daniel, Tim R. Sass, J.R. Lockwood, and Kata Mihaly (2009). "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy* 4(4): 572-606.

McCaffrey, Daniel, J.R. Lockwood, Daniel Koretz, and Laura Hamilton (2003). "Evaluating Value-Added Models for Teacher Accountability." Santa Monica, CA : Rand Corporation.

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.

Raudenbush, Stephen W. and Anthony S. Bryk (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417-58.

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

Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher (2010). *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*. Nashville, TN: National Center on Performance Incentives, Vanderbilt University.

Appendix B. Tables and Figures

Table 1: Randomization check

	(1) Baseline math	(2) Baseline reading	(3) Female	(4) Latino	(5) Black	(6) White	(7) F/R lunch	(8) Spec. ed.	(9) Neigh. pov. conc.	(10) Neigh. soc. stat.
Win	-0.1418	-0.0485	0.0831	0.0070	-0.0204	0.0089	0.0409	0.0587	0.0938	-0.0483
	(0.096)	(0.096)	(0.051)	(0.011)	(0.014)	(0.008)	(0.041)	(0.030)	(0.080)	(0.069)
N	381	381	381	381	381	381	381	381	381	381
Lott. mean	-0.0351	0.1476	0.5620	0.0105	0.9816	0.0052	0.8127	0.0919	0.5248	0.1534
Lott. S.D.	0.9201	0.9247	0.4955	0.1021	0.1345	0.0724	0.3897	0.2892	0.7728	0.6640
Syst. mean	0.0795	0.0967	0.4978	0.3545	0.5225	0.0901	0.8645	0.1321	0.2412	-0.3068
Syst. s.d.	0.9951	0.9732	0.5000	0.4784	0.4995	0.2863	0.3423	0.3386	0.8123	0.8240
Syst. N	144,197	144,197	144,197	144,197	144,197	144,197	144,197	144,197	144,197	144,197

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: first row displays coefficient and standard error from a regression of the pre-randomization variable on the winning indicator and lottery fixed effects.

Table 2: Number of winners and losers in the charter school lotteries

Year of Lottery	Grade	Lost	Won	Total in year-grade (VA)
2006	6th	45	32	24,660
2007	6th	11	39	24,523
2009	6th	17	88	23,797
2009	8th	32	21	24,256
2010	4th	29	6	21,586
2010	7th	23	38	24,225

Note: the last column represents the number of students in that year and grade used in the value-added model. The year is when the lottery took place, so the first post-lottery school year for the first row is 2006-07, etc.

Table 3: Math: lottery- and VA-based results

	(1) Lott-1	(2) Lott-2	(3) FE-P	(4) FE-G	(5) FE-Y	(6) FE	(7) AR-P	(8) AR-Y	(9) AR	(10) RE
Win	-0.0925	-0.0881	-0.0478	-0.0645	-0.0593	-0.0517***	-0.0319***	-0.0450***	-0.0599***	-0.0983*
	(0.056)	(0.057)	(N/E)	(N/E)	(N/E)	(0.020)	(0.006)	(0.010)	(0.022)	(0.044)
Baseline math	0.7721***	0.7743***								
	(0.038)	(0.038)								
Baseline read	0.0664	0.0512								
	(0.037)	(0.038)								
Demographic controls	N	Y								
constant	-0.0093	-0.0501	-0.0159*	-0.0140	-0.0332**	-0.0536***	-0.0166***	-0.0477***	-0.0599***	-0.0680*
	(0.040)	(0.069)	(0.006)	(0.009)	(0.013)	(0.010)	(0.004)	(0.008)	(0.012)	(0.030)
N	381	381	381	381	381	381	381	381	381	381
Difference			0.040	0.024	0.029	0.036	0.056	0.043	0.028	-0.010
S.E.						(0.051)	(0.056)	(0.054)	(0.052)	(0.060)
School S.D.			0.121	0.135	0.143	0.190	0.073	0.105	0.204	0.374
VA obs			600,273	600,273	600,273	143,047	600,273	600,273	143,047	143,047
VA schools			560	560	560	540	560	560	540	540

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: columns 1 and 2 contain lottery-based estimates, with column 2 including demographic controls. The rest of the columns contain VA-based estimates of the lottery effect. All columns include fixed effects for each lottery (i.e. the year-by-grade combination of the lottery entered).

The VA model used to generate the school effect estimates varies across columns 3 through 10 as follows: columns 3 through 6 are fixed effects models. Column 3 pools years and grade levels together, with year-by-grade fixed effects. Column 4 pools grades together (with grade fixed effects) while estimating each year separately; while column 5 pools years together (with year fixed effects) and estimates a separate model for each grade. Column 6 estimates each year-by-grade combination separately. Columns 3 and 4 do not use empirical Bayes estimates because the estimated "true" variance is negative (see text).

Column 7 is an average residual model following Chetty et al. (2011) and pooling years and grades with year and grade fixed effects. Column 8 is the same as 7 except it estimates each grade separately, with year fixed effects. The models in columns 9 and 10 both estimate each year-by-grade combination separately. Column 9 is an average residual model following Deming et al. (2011), and does not use an empirical Bayes estimate; while column 10 is the random effects model described in the text. The final rows represent the difference between the VA estimate of the ITT effect and the lottery estimate; the bootstrapped standard error of the difference; the estimated standard deviation of school effects of the VA model; and the number of observations and schools in the VA model.

Table 4: Reading: lottery- and VA-based results

	(1) Lott-1	(2) Lott-2	(3) FE-P	(4) FE-G	(5) FE-Y	(6) FE	(7) AR-P	(8) AR-Y	(9) AR	(10) RE
Win	-0.0199	-0.0194	-0.0133	-0.0112	-0.0276	-0.0180	-0.0118**	-0.0193**	-0.0245	-0.0387
	(0.064)	(0.065)	(N/E)	(N/E)	(N/E)	(0.019)	(0.005)	(0.007)	(0.022)	(0.038)
Baseline math	0.2919***	0.2876***								
	(0.044)	(0.043)								
Baseline read	0.5609***	0.5413***								
	(0.043)	(0.044)								
Demographic controls	N	Y								
constant	0.0343	0.0214	0.0043	0.0015	-0.0268**	-0.0068	-0.0095***	-0.0247***	-0.0033	0.0403
	(0.047)	(0.080)	(0.005)	(0.007)	(0.008)	(0.009)	(0.003)	(0.005)	(0.011)	(0.025)
N	381	381	381	381	381	381	381	381	381	381
Difference			0.006	0.008	-0.008	0.001	0.008	0.000	-0.005	-0.019
S.E.						(0.060)	(0.065)	(0.064)	(0.063)	(0.067)
School S.D.			0.120	0.124	0.142	0.142	0.050	0.068	0.162	0.337
VA obs			600,926	600,926	600,926	143,047	600,926	600,926	143,047	143,047
VA schools			560	560	560	540	560	560	540	540

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: columns 1 and 2 contain lottery-based estimates, with column 2 including demographic controls. The rest of the columns contain VA-based estimates of the lottery effect. All columns include fixed effects for each lottery (i.e. the year-by-grade combination of the lottery entered).

The VA model used to generate the school effect estimates varies across columns 3 through 10 as follows: columns 3 through 6 are fixed effects models. Column 3 pools years and grade levels together, with year-by-grade fixed effects. Column 4 pools grades together (with grade fixed effects) while estimating each year separately; while column 5 pools years together (with year fixed effects) and estimates a separate model for each grade. Column 6 estimates each year-by-grade combination separately. Columns 3, 4 and 5 do not use empirical Bayes estimates because the estimated "true" variance is negative (see text).

Column 7 is an average residual model following Chetty et al. (2011) and pooling years and grades with year and grade fixed effects. Column 8 is the same as 7 except it estimates each grade separately, with year fixed effects. The models in columns 9 and 10 both estimate each year-by-grade combination separately. Column 9 is an average residual model following Deming et al. (2011), and does not use empirical Bayes estimates; while column 10 is the random effects model described in the text. The final rows represent the difference between the VA estimate of the ITT effect and the lottery estimate; the bootstrapped standard error of the difference; the estimated standard deviation of school effects of the VA model; and the number of observations and schools in the VA model.