2011 SREE Conference Abstract Template

Abstract Title Page

Not included in page count.

Title: Evaluating Value-Added Methods of Estimating of Teacher Performance

Author(s): Cassandra M. Guarino, Mark D. Reckase, Jeffrey M. Wooldridge

Abstract Body

Background / Context: Accurate indicators of educational effectiveness are needed to advance national policy goals of raising student achievement and closing social/cultural based achievement gaps. If constructed and used appropriately, such indicators for both program evaluation and the evaluation of teacher and school performance could have a transformative effect on the nature and outcomes of teaching and learning. Measures based on value-added models of student achievement (VAMs) are gaining increasing acceptance among policymakers as an improvement over conventional indicators of performance. Much controversy exists, however, as to the best way to construct VAMs and to their optimal application. A plethora of methods has been developed (e.g., Sanders & Horn, 1994; Sanders et al., 1997; McCaffrey et al., 2004; Raudenbush, 2009), and studies that compare estimates derived from different models have found substantial variability across methods (McCaffrey et al., 2004). Concerns remain that our understanding of these models is as yet limited and that incentives built around them may cause more harm than good, with teachers' unions, in particular, reluctant to allow their constituents to be judged on the basis of measures that are potentially biased. The few studies that have attempted to validate VAMs have drawn different conclusions (e.g., Kane & Staiger, 2008; Rothstein, 2008)[*], and questions about the validity of VAMs linger.

Purpose / Objective / Research Question / Focus of Study: This paper is the first in a series of papers that aims to resolve controversies surrounding VAMs. Our research question is: How well do different estimators perform in estimating teacher effects in a commonly used VAM framework?

We focus our study on six estimators—most of which are commonly used in the research literature and in policy applications involving teacher effects. We answer our research questions by first outlining the assumptions that must be met for each estimator to have good statistical properties in the context of a fairly common theoretical framework. We then apply the estimators to the task of recovering teacher effects in simulated data. We apply the estimators to different types of data, where each data set is generated to violate or maintain specific assumptions in the attempt to mimic different types of student grouping and teacher assignment scenarios, and compare their performance.

Setting: We use simulated data and conduct the research at Michigan State University.

Population / Participants / Subjects: We simulate data on students and teachers in grades 3, 4, and 5 in a hypothetical school district.

Intervention / Program / Practice: There is no intervention used in our study.

Research Design: Our empirical investigations consist of a series of Monte Carlo simulations to evaluate the quality of various VAM estimation approaches. We use artificially generated data to investigate how well different estimators recover true effects under different scenarios. These scenarios correspond to data generating processes (DGPs) that vary the mechanisms used to assign students to teachers. To data generated from each DGP, we apply the set of estimators discussed in Section 3. We then compare the resulting estimates with the true underlying effects.

To isolate fundamental problems, we restrict the DGPs to a relatively narrow set of idealized conditions. We assume that test scores are perfect reflections of the sum total of a

[*] Kane and Staiger (2008) compare experimental VAM estimates for a subset of Los Angeles teachers with earlier non-experimental estimates for those same teachers and find that they are similar. Rothstein (2008) devises falsification tests that challenge the validity of VAM-based measures of teacher performance in North Carolina.

child's learning and that they are on an interval scale that remains constant across grades. We assume that there are no time-varying child or family effects, no interactions between students and teachers or schools, no peer effects, and we assume that unobserved child-specific heterogeneity has constant effect in each time period. We also assume that the GDL assumption holds—namely, that decay in schooling effects is constant over time. Finally, there are no time effects embedded in our DGPs. Thus we provide idealized conditions under which to test the performance of the estimators in uncovering teacher effects. If they fail under these conditions, they will do worse under the more complex conditions of the educational system.

To mirror the basic structural conditions of an elementary school system for, say, grades 3 through 5 over the course of three years, we create data sets that contain students nested within teachers nested within schools, with students followed longitudinally over time. Our simple baseline DGP is as follows:

$$\begin{aligned} A_{i3} &= \lambda A_{i2} + \delta_{i3} + \beta_{i3} + c_i + e_{i3} \\ A_{i4} &= \lambda A_{i3} + \delta_{i4} + \beta_{i4} + c_i + e_{i4} \qquad (13) \\ A_{i5} &= \lambda A_{i4} + \delta_{i5} + \beta_{i5} + c_i + e_{i5} \end{aligned}$$

where A_{i2} is a baseline score reflecting the subject-specific knowledge of children entering third grade, λ is a time constant decay parameter, c_i is a time-invariant child-specific growth effect, β_{it} is the teacher-specific contribution to growth (the true teacher value-added effect), δ_{it} is the school-specific contribution to growth (the true school effect), and e_{it} is a random deviation for each student. (Because we assume that A_{i3} depends on A_{i2} using the same decay λ, it makes sense to think of A_{i2} as a second-grade test score or a pre-test score.) Because we assume independence of e_{it} over time, we are maintaining the common factor restriction in the underlying cumulative effects model. We assume that the time-invariant child-specific growth effect c_i is uncorrelated with the baseline test score A_{i2}.

In the simulations reported in this paper, we do not include school effects, thus the δ_{it} are set to zero. The random variables A_{i2}, β_{it}, c_i, and e_{it} are drawn from normal distributions, where we adjust the standard deviations to allow different relative contributions to the scores. In the simulations we report, the standard deviation of the student-specific effects is 0.5, the standard deviation of the teacher effects is 0.25, and the standard deviation of the e_{it} is 1.0. In all cases, the specific values of the variables were randomly selected so the correlations among them was assumed to be 0. As a result, in the case where $\lambda = 1$, the teacher effect explains approximately $1/20^{th}$ of the variance of the gain score. We chose these parameters as a first approximation to generating data that would be consistent with data observed in state testing programs that have vertically scaled tests in place. In those testing programs, the effect size of student growth for one academic year of instruction ranges from .2 to .5. Because it is believed that few students actually decline in achievement after a year of instruction, the proportion of the combined change in performance due to all variables that is negative should be small. If the amount of growth on average is assumed to be .5 standard deviations, the standard deviation of student growth should be no larger than .5 so that 16% or less of the students will show negative growth. With the values for the generating model parameters given above, obtaining a standard deviation of student growth of less .5 or less requires a value for λ of less than 1. A value of .5 was used for half of the simulations. There is also the assumption that the teacher effect is less than the student effect. We chose the value for the standard deviation of e_{it} to be 1, indicating that it was responsible for more of the variation in test scores than the teacher and student fixed effects. However, we tested the sensitivity of our results to other parameterization choices, and they were relatively robust.

Our data structure has the following characteristics that do not vary across simulations:

- 10 schools
- 3 grades (3^{rd}, 4^{th}, and 5^{th}) of scores and teacher assignments, with a base score in 2^{nd} grade
- 4 teachers per grade
- 20 students per classroom
- 4 cohorts of students
- No crossover of students to other schools

To create different scenarios, we vary certain key features: the sorting of students and teachers into schools, the grouping of students into classes, the assignment of classes of students to teachers within schools, and the amount of decay in prior learning from one period to the next. Within each of the school-sorting cases, we study the 10 different mechanisms for the assignment of students outlined in Table 1. (Insert Table 1 here) Finally, we vary the decay parameter λ as follows: (1) $\lambda = 1$ (no decay or complete persistence) and (2) $\lambda = .5$ (fairly strong decay). The DGPs chosen for each simulation reproduce scenarios in which the assumptions mentioned above either hold or are violated. Thus, we explore $3 \times 10 \times 2 = 60$ different scenarios in this paper. In the future we will allow school effects; for now, our simulations set the school effects to zero. We use 500 Monte Carlo replications per scenario in evaluating each estimator.

Analysis: The following are the estimating equations we use, reflecting the simplifications determined by our DGPS. Specifically, we remove the time-varying intercept because our data have no time effects, we have no time-varying child and family effects, and we assume that $\pi_t = 1$:

$$\Delta A_{it} = E_{it}\beta_0 + c_i + e_{it} \qquad (14)$$

$$A_{it} = \lambda A_{i,t-1} + E_{it}\beta_0 + c_i + e_{it} \qquad (15)$$

$$\Delta A_{it} = \lambda \Delta A_{i,t-1} + \Delta E_{it}\beta_0 + \Delta e_{it} \qquad (16)$$

where E_{it} is the vector of teacher dummies.

For each of the 500 iterations pertaining to one DGP, we estimate effects for each of the 120 teachers using one of six estimation methods: pooled OLS (POLS) applied to (14) – so that the presence of c_i is effectively ignored in estimation – random effects (RE) applied to (14), fixed effects (FE) applied to (14), POLS applied to (15) (which we have called dynamic OLS or DOLS), instrumental variables (a simplified version of Arellano and Bond that we call AB for "first-difference instrumental variables") applied to (16), and Blundell and Bond (BB) applied to (16).

Evaluating the Estimators

For each iteration and for each of the six estimators, we save the estimated individual teacher effects, which are the coefficients on the teacher dummies and collapse these data to the teacher level retaining the true teacher effects generated from the simulation in the data as well. To study how well the methods uncover the true teacher effects, we adopt some simple summary measures. We regress the estimated 119 teacher effects (where one teacher is the base case) on the true effects generated from the simulation. From this simple regression, we report the average coefficient and its standard deviation across the 500 simulations. Regressing the estimated teacher effects on the true teacher effects tells us whether the estimated teacher effects are correct when compared with the average teacher. To see this, recall that we can write a simple regression equation as

$$\hat{\beta}_j - \hat{\beta} = \hat{\theta}(\beta_j - \bar{\beta}) + residual_j \qquad (17)$$

where $\hat{\beta}_j$ is the estimated effect of teacher *j* (obtained using a particular estimation approach), β_j is the true effect of teacher *j*, the overbars represent averages (across the 119 teachers), and $\hat{\theta}$ is the simple regression coefficient. If $\hat{\theta} = 1$ then a movement of β_j away from its mean is tracked by the same movement of the estimate away from *its* mean. In effect, $\hat{\beta}_j - \hat{\bar{\beta}}$ is an unbiased estimator of $\beta_j - \bar{\beta}$ for all $j = 1, \ldots, 119$. Notice that this is different from saying $\hat{\beta}_j$ is unbiased for β_j, something we cannot guarantee because the means of the $\hat{\beta}_j$ and the β_j will typically differ. (In effect, the mean value of the β_j is not identified, so we can only ask how well the estimators are recovering the teacher effects compared with the mean.)

If $\hat{\theta} > 1$ then the estimation procedure is magnifying the differences of teachers relative to the mean. For example, a teacher that is slightly above average can look much better than the average because $\beta_j - \bar{\beta}$ is multiplied by $\hat{\theta}$. Similarly, a teacher just below average can appear much worse than average. If $\hat{\theta} < 1$ (but positive) then the estimation method shrinks the estimated effects closer to its mean, making it more difficult than it should be to distinguish any teacher from the average. Because there is always sampling error in the estimates, we report the average of the $\hat{\theta}$, which we denote $\bar{\hat{\theta}}$, across the 500 replications.

The standard deviation of the $\hat{\theta}$ across simulations tells us how much the estimate deviates from its average. So, if the average value, $\bar{\hat{\theta}}$, is near one and the standard deviation is small, then much of the time the estimated effects are properly tracking the true effect. A larger standard deviation means the estimates are correct on average but variable.

As we already discussed, if $\bar{\hat{\theta}}$ is very far from one and the standard deviation is small, then the estimated effects are systematically magnifying or compressing the teacher effects (relative to the mean, as always). Coupled with a small standard deviation we can conclude that the estimated teacher effects are far off in most simulations. This is an undesirable situation.

We also report the average Spearman rank correlation between the estimated and true effects across the 500 iterations. We do so because, as we mentioned, certain policy applications of value-added measures of teacher performance may rely solely on rankings and not on the magnitude of the deviation from the average. Thus the rank correlation tells us how well the estimators preserve the true rankings.

Findings / Results: Our findings for the case in which both students and teachers are randomly sorted across schools and when λ=1 are shown in Tables 2A and 2B (full findings not reported in this abstract). All estimators work well when students are randomly grouped in classrooms and classrooms are randomly assigned to teachers. Our main finding is that no one method is guaranteed to reliably capture true teacher effects in all contexts, although some are more robust than others. Because we consider a variety of DGPs, student grouping mechanisms, and teacher assignment mechanisms, it is not surprising that no single method works well in all contexts. That the POLS and RE estimators work well under random assignment (regardless of how students are grouped), even when λ is much less than unity, suggests that the estimators are fairly robust in such situations with only misspecified dynamics. Unfortunately, misspecified dynamics leads to much worse performance when assignment is based on past performance or on the unobserved student effect. Thus, these estimators cannot be trusted for certain realistic assignment mechanisms. Instead, estimating a simple dynamic regression by OLS was superior – often by a large margin – in situations of dynamic grouping of students with nonrandom assignment to teachers.

The FE estimator is fairly robust to dynamic misspecification under random assignment and to either form of static assignment. The downside is that dynamic misspecification in conjunction with dynamic assignment cause FE to be very badly behaved: the rank correlation between the estimated and true teacher effects is actually negative and almost .5 in magnitude in the case with the most decay. Again, the DOLS estimator works much better. One caveat is that our simulations use a fairly small variance for the student effect in the gain-score equation. The evidence suggests DOLS will work less well as student heterogeneity becomes relatively more important.

The simulations show that the Arellano and Bond approach can be an attractive alternative to FE, especially when FE is subject to dynamic misspecification. In fact, except in two cases of dynamic grouping of students – DG-RA and DG-NA – AB outperforms FE when $\lambda < 1$. However, it is clearly inferior to DOLS in cases of dynamic assignment. Overall, however, the performance of estimators that attempt to eliminate heterogeneity suffers in contexts in which teachers are non-randomly sorted across schools; these estimators are no longer reliable.

Our findings suggest that choosing estimators on the basis of structural modeling considerations may produce inferior results. The DOLS estimator is never the prescribed approach under the structural cumulative effects model with a geometric distributed lag (unless there is no student growth heterogeneity), yet it is often the best estimator. That is not to say that the general cumulative effects model is incorrect. It merely reflects the fact that the assumptions needed to make it tractable – linearity, the geometric distributed lag, the common factor restriction – along with the possibility of nonrandom assignment may yield estimators that are poorly behaved. The findings in this paper, though special, suggest that flexible approaches based on dynamic treatment effects (for example, Lechner (2008), Wooldridge (2010, Chapter 21)) may be fruitful: one can think of the DOLS estimator as a regression-based version of a dynamic treatment effects estimator.

Conclusions: This study has taken the first step in evaluating different value-added estimation strategies in the conditions under which they are most likely to succeed. It is clear from this study that many VAMs hold promise: they may be capable of overcoming obstacles presented by non-random assignment and provide valuable information. Given their context-dependency, however, caution must be applied in interpreting the findings that VAMs, as currently applied in practice and in the research literature, are producing. They may best be viewed as suggestive evidence of effects, at this point. In addition, methods of constructing estimates of teacher effects that we can trust for high-stakes evaluative purposes must be further studied. Clearly, although value-added measures of teacher performance hold promise, before they can be used for policy purposes, more research is needed. There is much left to investigate. An important issue to study is the effect of failure of the common factor restriction on the various procedures. The DGP that we used would be much more flexible if we allow the errors in the cumulative effects formulation to be an unrestricted AR(1) process. As mentioned earlier, with more general DGPs we should add dynamic treatment effects estimators to the list of competitors. In addition, if contextual problems—grouping and assignment mechanisms—can be deduced from available data, then it may be possible to determine which estimators should be applied in a given context. For this purpose, structural modeling considerations may be helpful in that they yield tests that have the potential to identify violations of particular assumptions. In addition to the relaxation of assumptions and the development of contextual diagnostics, further research is needed regarding the ability of more detailed astructural approaches to detect treatment effects with greater accuracy. Investigations in these areas are the subject of current research in progress by the authors.

Appendices

Appendix A. References

Arellano, M. & Bond, S. (1991) Some Tests of Specification of Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*, 58, pp.277-298.

Blundell, R. & Bond, S. (1998) Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics*, 87, 11-143.

Briggs, D. &Weeks, J. (2009) The Sensitivity of Value-Added Modeling to the Creation of a Vertical Score Scale, *Education Finance and Policy,* 4(4), 384-414.

Hanushek, E. "The Economics of Schooling: Production and Efficiency in the Public Schools," *Journal of Economic Literature*, XXIV (3): 1141-78, 1986.

Hanushek, E. "Conceptual and Empirical Issues in the Estimation of Educational Production Functions," *Journal of Human Resources*, 14(3): 351-388, 1979.

Harris, D. & Sass, T. (2006) Value-Added Models and the Measurement of Teacher Quality, Unpublished Draft.

Kane, T. & Staiger, D. (2008) Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation, Working Paper 14607, National Bureau of Economic Research.

Lechner, M. (2008), Matching Estimation of Dynamic Treatment Models: Some Practical Issues, in *Advances in Econometrics*, Volume 21 (Modeling and Evaluating Treatment Effects in Econometrics). Daniel Millimet, Jeffrey Smith, and Edward Vytlacil (eds.), 289-333-117. Amsterdam: Elsevier, 2008

Martineau, J. (2006) Distorting Value Added: The Use of Longitudinal, Vertically Scaled Student Achievement Data for Growth-Based, Value-Added Accountability, *Journal of Educational and Behavioral Statistics*, 31(1), pp. 35-62.

McCaffrey, D., Lockwood, J.R., Louis, T., & Hamilton, L. (2004) Models for Value-Added Models of Teacher Effects. *Journal of Educational and Behavioral Statistics*, 29(1), pp. 67-101.

Raudenbush, S. (2009) Adaptive Centering with Random Effects: An Alternative to the Fixed Effects Model for Studying Time-Varying Treatments in School Settings. *Education Finance and Policy,* 4(4), 468-491.

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. Applied Psychological Measurement, 9, 401–412.

Reckase, M. D. & Li, T. (2007). Estimating gain in achievement when content specifications change: a multidimensional item response theory approach. In R. W. Lissitz (Ed.) Assessing and modeling cognitive development in school. Maple Grove, MN: JAM Press.

Rothstein, J. (2008) Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *NBER Working Paper Series,* Working Paper 14442, http://www.nber.org/papers/w14442.

Sanders, W. & Horn, S. (1994) The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment. *Journal of Personnel in Education*, 8, 299-311.

Sanders, W., Saxton, A., & Horn, B. (1997) The Tennessee Value-Added Assessment System: A Quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), Grading Teachers, Grading Schools: Is student Achievement a Valid Evaluational Measure? Thousand Oaks, CA: Corwin Press, Inc., 137-162.

Todd, P. & Wolpin, K. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *Economic Journal*, 113(485), 3-33.

US Department of Education (2009) Race to the Top Program: Executive Summary, http://www2.ed.gov/programs/racetothetop/executive-summary.pdf, accessed on 9/8/10.

Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2e. MIT Press: Cambridge, MA.

Zeger, S., Liang, K., & Albert, P. (1988) Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*. 44(4), 1049-1060.

Appendix B. Tables and Figures

Table 1: Grouping and Assignment Acronyms

Acronym	Process for grouping students in classrooms	Process for assigning students to teachers
RG-RA	Random	Random
DG-RA	Dynamic (based on prior test scores)	Random
DG-PA	Dynamic (based on prior test scores)	Positive correlation between teacher effects and prior student scores (better teachers with better students)
DG-NA	Dynamic (based on prior test scores)	Negative correlation between teacher effects and prior student scores
BG-RA	Static based on baseline test scores	Random
BG-PA	Static based on baseline test scores	Positive correlation between teacher effects and baseline student scores
BG-NA	Static based on baseline test scores	Negative correlation between teacher effects and baseline student scores
HG-RA	Static based on heterogeneity	Random
HG-PA	Static based on heterogeneity	Positive correlation between teacher effects and student fixed effects
HG-NA	Static based on heterogeneity	Negative correlation between teacher effects and student fixed effects

Table 2A: Average estimates of θ and λ over 500 replications in the teacher-level regressions of estimated effects on true effects: Case 1, λ=1

Estimator / Assignment Mechanism	POLS	DOLS	RE	FE	FDIV	BB
RG-RA	1.000 (0.047)	1.000 (0.053) λ=1.104	1.001 (0.045)	0.997 (0.096)	0.959 (0.103) λ=.906	1.079 (0.124) λ=1.185
DG-RA	1.004 (0.068)	1.002 (0.051) λ=1.088	1.005 (0.060)	0.895 (0.193)	0.689 (0.216) λ=.363	1.066 (0.413) λ=1.162
DG-PA	1.343 (0.058)	1.000 (0.068) λ=1.107	1.266 (0.056)	-0.410 (0.135)	0.480 (0.211) λ=.131	-1.561 (0.297) λ=1.648
DG-NA	0.638 (0.057)	0.993 (0.068) λ=1.096	0.700 (0.055)	2.086 (0.157)	0.540 (0.257) λ=.269	3.493 (0.338) λ=1.630
BG-RA	1.002 (0.047)	1.002 (0.063) λ=1.188	1.002 (0.045)	0.995 (0.096)	0.955 (0.103) λ=.897	1.076 (0.126) λ=1.169
BG-PA	1.003 (0.049)	0.733 (0.061) λ=1.135	1.002 (0.048)	0.988 (0.102)	0.950 (0.111) λ=.895	1.062 (0.127) λ=1.201
BG-NA	0.996 (0.048)	1.233 (0.065) λ=1.126	0.997 (0.046)	0.988 (0.110)	0.949 (0.117) λ=.903	1.082 (0.144) λ=1.185
HG-RA	1.003 (0.090)	1.002 (0.087) λ=1.099	1.003 (0.075)	0.999 (0.097)	0.970 (0.101) λ=.953	1.080 (0.126) λ=1.199
HG-PA	1.641 (0.068)	1.567 (0.070) λ=1.080	1.506 (0.061)	0.995 (0.094)	0.977 (0.097) λ=.853	1.083 (0.131) λ=1.177
HG-NA	0.360 (0.068)	0.402 (0.067) λ=1.090	0.495 (0.060)	0.996 (0.100)	0.935 (0.116) λ=.930	1.074 (0.124) λ=1.189

Table 2B: Average Spearman rank correlation over 500 replications between the estimated teacher effects and the true teacher effects: Case 1, λ =1

Estimator / Assignment Mechanism	POLS	DOLS	RE	FE	FDIV	BB
RG-RA	0.883 (0.025)	0.842 (0.031)	0.889 (0.024)	0.671 (0.082)	0.638 (0.079)	0.592 (0.081)
DG-RA	0.782 (0.039)	0.837 (0.031)	0.816 (0.036)	0.378 (0.085)	0.255 (0.078)	0.229 (0.089)
DG-PA	0.896 (0.018)	0.842 (0.033)	0.902 (0.018)	-0.307 (0.083)	0.188 (0.083)	-0.423 (0.066)
DG-NA	0.640 (0.063)	0.827 (0.034)	0.709 (0.058)	0.779 (0.048)	0.199 (0.084)	0.693 (0.053)
BG-RA	0.884 (0.024)	0.795 (0.037)	0.889 (0.023)	0.665 (0.082)	0.633 (0.081)	0.590 (0.082)
BG-PA	0.885 (0.023)	0.705 (0.056)	0.890 (0.022)	0.669 (0.081)	0.638 (0.082)	0.592 (0.082)
BG-NA	0.883 (0.025)	0.866 (0.024)	0.887 (0.024)	0.662 (0.084)	0.631 (0.082)	0.577 (0.083)
HG-RA	0.694 (0.049)	0.702 (0.048)	0.755 (0.042)	0.672 (0.082)	0.647 (0.080)	0.588 (0.083)
HG-PA	0.914 (0.015)	0.900 (0.017)	0.918 (0.014)	0.670 (0.079)	0.658 (0.075)	0.560 (0.080)
HG-NA	0.387 (0.087)	0.420 (0.082)	0.556 (0.074)	0.668 (0.078)	0.602 (0.084)	0.610 (0.076)