Journal of Educational and Behavioral Statistics

What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice?

Stephen W. Raudenbush

The online version of this article can be found at:

Published on behalf of

American Educational Research Association

and

⑤SAGE

http://www.sagepublications.com

Additional services and information for *Journal of Educational and Behavioral Statistics* can be found at:

Email Alerts: http://jebs.aera.net/alerts

Subscriptions: http://jebs.aera.net/subscriptions

Reprints: http://www.aera.net/reprints

Permissions: http://www.aera.net/permissions

Citations: http://jeb.sagepub.com/content/29/1/121.refs.html

>> Version of Record - Jan 1, 2004

What is This?

What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice?

Stephen W. Raudenbush
University of Michigan

The question of how to estimate school and teacher contributions to student learning is fundamental to educational policy and practice, and the three thoughtful articles in this issue represent a major advance. The current level of public confusion about these issues is so severe and the consequences for schooling so great that it is a big relief to see this journal highlight the key issues.

A common theme in these articles is that we should compare schools or teachers by comparing their "value added" to student learning rather than by comparing unadjusted mean levels of achievement or, as is currently common practice, the percent of students in a school or class who are classified as "proficient." As Ballou, Sanders, and Wright (BSW) note, it makes no sense to hold schools accountable for mean achievement levels when students enter those schools with large mean differences in achievement. Moreover, given the remarkable mobility of students across schools, particularly in large urban districts, changes in mean achievement at the school level may bear little relation to instructional effectiveness.

In contrast, the value-added philosophy is to hold schools and teachers accountable for the learning gains of students they serve. This seems simple enough, yet the technical questions raised in these articles are many: whether and how to adjust for covariates, whether teachers (or schools) should be treated as fixed or random, how to represent cumulative effects of teachers or schools, how to model covariation in student responses and teacher effects, whether and how to incorporate multiple cohorts, and how to formulate models that appropriately handle missing data.

A prior question is: "What are we trying to estimate with these models?" School and teacher effects are causal effects, yet the treatments students experience and the potential outcomes under alternative treatments (Rubin, 1978; Rosenbaum & Rubin, 1983; Holland, 1986) are not clearly defined in these discussions. As a result, we are not clear about the experiments we are trying to approximate with value-added analyses or, therefore, about the prospects of achieving reasonable approximations. In my view, defining possible treatments and potential outcomes eliminates some of the confusion by showing what kinds of effects can and cannot reasonably be estimated.

Two Kinds of Effects

Raudenbush and Willms (1995) (RW) defined two kinds of causal effects that might be estimated in a school accountability system. The first or "Type A" effect

121

is of interest to a parent selecting a school for her children. The second, or "Type B" effect is of interest to district or state administrators who wish to hold school personnel accountable for their contributions to student outcomes. RW described plausible conditions for unbiased estimation of Type A effects. In contrast, they found the prospects for Type B effects unpromising given the kind of data available in accountability systems.[1]

RW reasoned that the child's potential outcomes would be a function of pre-assignment student characteristics *S,* random error *e,* and two aspects of schools: school context, *C,* and school practice, *P*. *C* includes the social environment of the school (e.g., the neighborhood in which it is located) and the social composition of the school. Teachers and administrators have little or no control over *C,* though *C* might strongly contribute to school effectiveness through peer interactions, parent involvement, social norms, and the availability of role models (Coleman et al., 1966; Willms, 1986; Lee & Bryk, 1989). In contrast, school leaders and teachers do have substantial influence over *P,* though *P* is likely also associated with *C.*

Type A Effect

In terms of the Rubin causal model, the Type A effect (of interest to parents) is the difference between child *i*'s potential outcome in school *j,* say $Y_{ij}(S_i,C_j,P_j,e_{ij})$ and that child's potential outcomes in school *j*′, that is, $Y_{ij'}(S_i,C_{j'},P_{j'},e_{ij'})$. RW reasoned that parents would be indifferent regarding the relative contributions of *C* and *P* to this effect. Therefore, an experiment that would reveal the Type A effect for parent *i* would be a study in which students having a common $S = S_i$ were randomly assigned to either school *j* or school *j*′.[2] Treatment assignment would be ignorable (independent of *S*) and so the expected treatment effect estimate for comparing schools *j* and *j*′ would depend only on $C_j,P_j,C_{j'},P_{j'}$. Without the benefit of randomization, one might obtain an unbiased estimate of the same causal effect by controlling for observed student-level covariates *X* under the assumption of strong ignorability, namely that the potential outcomes are not associated with school assignment after controlling for *X*. In particular, this assumption implies that *X* captures the association between *S* and school assignment, so that only $C_j,P_j,C_{j'},P_{j'}$ contribute systematically to the estimated school effects. Type A effects are arguably estimable with tolerably small bias because the data available to school accountability analysts include some *X*s that likely are extremely important in explaining the link between student background and school assignment. In particular, schools that collect historical data on student achievement along with ethnicity and poverty status provide *X*s that are likely very informative about potential outcomes.

Type B Effect

In contrast, the Type B effect (of interest to district or state officials) is the difference between child *i*'s potential outcome in school *j* when school practice P_j^* is in operation, yielding $Y_{ij}^*(S_i,C_j,P_j^*,e_{ij}^*)$ and that child's potential outcomes in school *j* when school practice P_j, is in operation that is, yielding $Y_{ij}(S_i,C_j,P_j,e_{ij})$. RW reasoned that district or state officials would not want to hold school personnel

122

accountable for *C,* over which those personnel have no control. Officials would, however, want to hold personnel accountable for their practice, *P*. Importantly, the accountability system, if effective, would lead to a change in *P,* but not, at least in the short run, to a change in *C*.[3] Therefore, an experiment that would reveal the Type B effect would be a study in which *schools* were assigned at random to *practices, P*.[4] Treatment assignment would be ignorable (independent of *S* and *C*), and so the expected treatment effect estimate would depend only on $P^*_{j'}$ and P_j. Without the benefit of randomization, one might obtain an unbiased estimate of the same causal effect by controlling for observed student-level covariates, *X,* and school-level covariates, *W,* under the assumption of strong ignorability, namely that the potential outcomes are not associated with the school-level treatment assignment after controlling for *X* and *W*. Thus, strong ignorability implies that *X,W* capture the association between *S,C* and the assignment of schools to *P*.

The problem with non-experimental approximations to the school-based randomized trial is not that covariates *X,W* are unavailable. The difficulty is that school practice *P* is not defined, much less observed! Therefore we cannot assess which *X*s and *W*s are correlated to treatment assignment. A common practice in school accountability research is to regress the outcome on *X* and *W* and to assume that the school mean residual is a good estimate of *P*. But this practice cannot reveal the effect of *P* unless we assume that *P* is uncorrelated with *X* and *W*.[5] Thus, the prospects for estimating Type B effects are dim at best.

Implications for Value Added Models (VAM)

This reasoning concerning Type A and Type B effects has important implications for VAM. I believe it explains in part why BSW expressed discomfort interpreting VAM results when school-level poverty (as indexed by percent of students receiving free lunch) was controlled. It also helps explain why Tekwe et al. expressed uncertainty about their results controlling for covariates at both levels. And it explains in a conceptual way a vexing problem revealed in McCafferty et al.'s technical analysis: namely, that estimation of teacher effects is most problematic when schools serve very different kinds of students. Clearly, the more variable C_j is in the RW model, the more problematic it is to assume that VAM estimates correspond to P_j, the implicit object of interest in these articles.

The problem of estimating Type B effects is even more pronounced when the aim is to estimate school and teacher effects simultaneously, as in the VAM proposed by McCafferty et al. Classrooms as well as schools will be characterized by contextual conditions and practices that contribute to student learning independent of student background. A Type B analysis would aim to separate the effects of the practice at the school level and at the teacher level. Since practice is unobserved at both levels in accountability systems, this separation appears inaccessible in accountability analyses. However, the indeterminacy of school versus teacher effects is a nonissue for Type A effects. In this case the parent might first select a school and then a teacher within a school. Alternatively, the parent can look across all schools and classrooms and pick the classroom that has the highest expected value for his child, regardless

123

of whether that value is attributable to school context or practice or classroom context or practice. If we view the type A effect for a class to be the combined result of school and classroom context and practice, this effect can be estimated without bias conditional on the strong ignorability assumption that student-level covariates *X* account for the association between potential outcomes and classroom assignment.

As BSW point out, care must be taken in estimating and adjusting for *X* in estimating what I am calling Type A effects. They use a two-step procedure: estimate a regression using *X* as covariate with fixed effects of teachers. The coefficients for *X* are then estimates of the pooled, within-school coefficient, often denoted β_w. As RW point out, this estimation can easily be accomplished by centering *X* within teachers, obviating the need to enter teacher dummy variables. In the second step, an adjusted dependent $Y - X\hat{\beta}_w$ is used in the accountability analysis.

Modeling Teacher and School Effects on Student Growth

As the previous discussion shows, it does not appear possible to separate teacher and school effects using currently available accountability data. At one extreme, one might attribute all variation between classrooms to teachers. In that view, mean differences between schools are just differences in aggregate teacher effects. At the other extreme, all variation between schools is attributable to variation in the skill of school management and other school organizational features, including instructional coordination across grades, teacher collaboration, teacher control, and school-level resources. In this view, teachers can be held accountable only for the classroom variation within schools. A range of views are located on the continuum between these two extremes, but these views cannot be adjudicated without a theory of what makes schools and teachers effective and without a research agenda that explicitly assesses the causal effects at each level. In short, one needs good estimates of Type B effects at each level, but these are inaccessible at either level if the relevant school and classroom processes are not observed.

So VAM are best aimed at assessing the Type A effect defined as the *combined effects* of context and practice at the classroom and school levels. I believe it is useful to define the potential outcomes associated with this effect as a way of informing model specification, evaluation, and interpretation. A useful way to do so is to view each student as possessing a smooth trajectory that would describe that student's growth if that student encountered "average" teachers and schools. The Type A effect in any year is then defined as a deflection from this expected curve. Of course this assumes an equated metric over time, as these articles emphasize.

This idea is displayed in Figure 1. The dashed line describes a hypothetical student's expected trajectory given "average" schools and classrooms. This student encounters a "non-average" classroom (classroom *j*) at time *t,* yielding observed achievement $Y_{t+1}^{(j)}$ at time $t + 1$. If this student had instead encountered an average classroom at time *t,* the outcome would have been the counterfactual $Y_{t+1}^{(0)}$. The causal effect associated with attendance in classroom *j* is then $Y_{t+1}^{(j)} - Y_{t+1}^{(0)}$. This seems straightforward enough. But what about the causal effect of teacher *j′,* whom our student experiences at time $t + 1$? Presumably $Y_{t+2}^{(j')} - Y_{t+2}^{(0)}$ is the combined
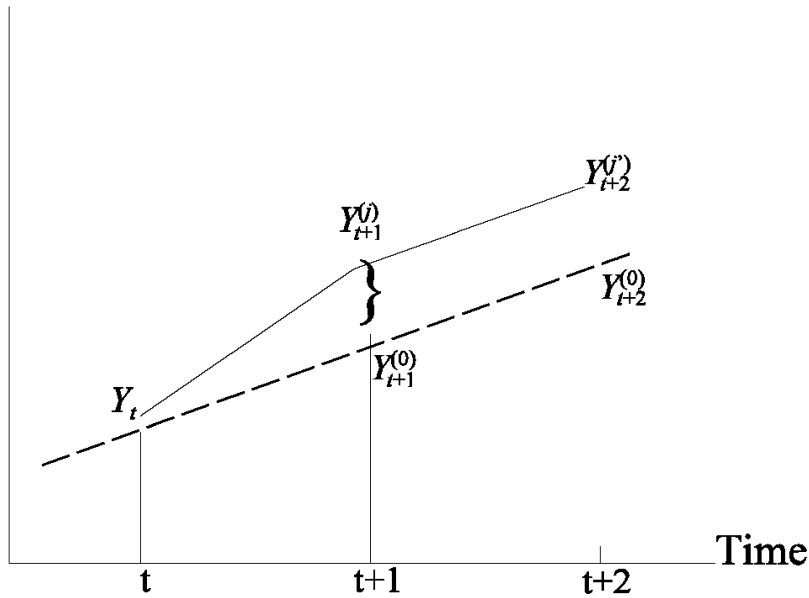
124

Outcome

FIGURE 1. *The dashed curve is the expected trajectory of a given student "average" for schools and teachers. For simplicity this student is "on trajectory" until time* t. *If assigned to teacher* j, *the student will exhibit outcome* $Y^{(j)}_{t+1}$. *The causal effect of teacher* j *is thus the deflection* $Y^{(j)}_{t+1} - Y^{(0)}_{t+1}$.

causal effect of having experienced teachers *j* and *j′*, but how should we decompose this combined effect into pieces attributable to the two teachers? McCafferty et al. make an extremely useful contribution by parameterizing a "rate of decay" in teacher effects over time. This enables the data to drive the decomposition rather than assuming *a priori* that effects are cumulative and additive.

A Polynomial Growth Model

To represent the conception of Figure 1 in the VAM, it seems sensible to represent each student's counterfactual expected trajectory as a polynomial of appropriate degree. This implies a random coefficient model for student growth augmented by a "deflection model" for value added (Raudenbush & Bryk 2002). In contrast, BSW use an unstructured covariance matrix to represent student contributions to the covariance structure with added random effects of teachers. And McCaffrey et al. express a preference for the unstructured covariance structure as more general than the random coefficient model illustrated in Raudenbush and Bryk or "RB." RB's illustrative example involved a polynomial of degree 1 or "straight-line" growth model. McCaffrey criticize such a model for placing strong restrictions on the vari-

125

ance structure (the model implies increasing variance if the correlation between intercept and slope is positive). Yet RB never recommended a life-long commitment to the straight-line model! In reality, the polynomial approach allows a range of models varying from simple (e.g., the straight-line model) to complex. Indeed, if the number of time points is *T*, then a $T - 2$ degree polynomial with time-specific within-subject variances is a saturated model identical to the unstructured model. A good argument can be made for selecting the lowest-order polynomial that reasonably fits the data. One may anticipate that the simpler model, if justified, will supply more precision in estimating teacher effects. It also is more flexible than the unstructured covariance matrix in allowing for the timing of testing to vary across students.

Consider a simple model for student growth and value added:

$$Y_i = A_i \pi_i + Z_i b + e_i, \tag{1}$$

where Y_i is a T_i by 1 vector of outcomes for student $i = 1, \ldots, n$, π_i is $p + 1$ vector of random coefficients, A_i is a known T_i by $p + 1$ design matrix with columns containing polynomial coefficients of degree *p,* and e_i a within subject error vector assumed for simplicity here to be distributed as $N(0, \sigma^2 I_{Ti})$. By design each student should have *T* observations but in fact only T_i outcomes were observed. Now Z_i is a T_i by *J* matrix having entries of 0 or 1 indicating whether student *i* had ever encountered teacher *j* by time $t = 1, \ldots, T_i$, and *b* is a *J* by 1 vector of teacher effects associated with teachers $j = 1, \ldots, J$ and assumed $N(0, \delta^2 I_j)$. For simplicity I omit covariates and assume $\pi_i \sim iid\ N(A_i \gamma, \tau)$. Note that I have assumed additive and cumulative teacher effects. However, I do so for simplicity of exposition here and acknowledge McCaffrey et al.'s advice to check and if necessary revise this assumption.

Then, given knowledge of the variance components and γ, the posterior mean of the teacher effects is given by

$$E(b|Y) = \left[\sum_{i=1}^{n} Z_i^T \left(I - A_i C_i^{-1} A_i^T \right) Z_i + \sigma^2 / \delta^2 I \right]^{-1}$$
$$\times \sum_{i=1}^{n} Z_i^T \left(Y - A_i \pi_i^* \right). \tag{2}$$

Here $\pi_i^* = A_i \gamma + C_i^{-1} A_i^T (Y_i - A_i \gamma)$ is the posterior mean of π_i and $C^{-1} = \sigma^2 (A_i^T A_i + \sigma^2 \tau^{-1})^{-1}$ is the posterior variance of π_i in a model without teacher effects.

Hence, Equation 2 represents a regression in which the outcome is, $Y_i - A_i \pi_i^*$, the discrepancy between the observed Y_i and its predicted value using the standard "empirical Bayes" polynomial coefficients. This outcome corresponds conceptually to the causal effect described in Figure 1 where the dashed curve is the empirical Bayes estimated polynomial for student *i.* If the left-hand side of Equation 2

were simply $\left(\sum_{i=1}^{n} Z_i^T Z_i \right)^{-1}$, these residuals would simply be averaged over the students taught by each teacher. The left-hand side would involve the inverse of a J by J matrix $\sum_{i=1}^{n} Z_i^T Z_i$, and would likely be ill-conditioned. The addition of the term $\sigma^2/\delta^2 I$ adds prior information, increasing precision through appropriate shrinkage and insuring the invertability of the matrix. Inclusion of the term $I - A_i C_i^{-1} A_i^T$ weights down students whose counterfactuals are estimated with large posterior variance as a result of missing data.

Here is an interesting trade off between assumptions and robustness that deserves more research. In Model 2 the variation between students is $Var(Y_i|b) = \Sigma_i = A_i \tau A_i^T + \sigma^2 I$. This is a stronger assumption than allowing an unstructured covariance matrix, Σ_i. However, if justified, this stronger assumption may make better use of the observed information, reducing the fraction of missing information and thereby increasing robustness to non-ignorable missingness while also increasing precision.

BSW wisely exploit the availability of tests in multiple subjects to improve the precision of estimation of teacher effects on any specific subject, and McCaffrey et al. include this multivariate approach in their general model. This multivariate outcome approach not only reduces confounding of teacher assignment with student background, as BSW indicate. It should also increase robustness of results to non-ignorable missingness.

Multiple Cohorts

Raudenbush, Bryk, and Ponisciak (2003) analyzed data collected on five cohorts of students over five years in Washington, DC. Even with over 50,000 students, precision in estimating teacher effects was modest. Using multiple cohorts appears essential to obtain adequate precision. Moreover, school effects were somewhat unstable, implying a need to average school effects over multiple cohorts in order to obtain a stable average effect. Finally, trends in improvement (gains in value added) cannot be estimated without multiple cohorts.

Fixed vs. Random Effects

Tekwe et al. find that a simpler fixed effect model produces similar "value added" effects than a more complex random effects model. However their interest is confined to estimating school effects with large samples of students and data with two time points. It is well known that the fixed effects and random effects estimates converge as cluster sizes grow large. Large cluster sizes do not apply, however, when teacher effects are of interest. And fixed effects models become unwieldy when multiple time points and multiple cohorts are available. Give that fixed effect estimates have good properties only in special circumstances, I would recommend random effects as a general approach.

127

Summary

In sum, the potential benefits of specifying low-order polynomial models can be combined with the benefits of multiple subject-area tests to yield a model with multiple growth curves per child. Covariates *X* included as indicated by BSW would add further information.

Such an approach may be useful in reducing confounding and increasing robustness to nonignorable missingness and is worthy of further research. Moreover, multiple cohorts can increase precision and allow study of change in valued added.

However, we must keep in mind that our estimates are, at best, Type A effects, of interest to parents selecting schools, not Type B effects, of interest to officials holding schools and teachers accountable for instructional practice. Certainly the estimates from VAM, when combined with other information, have potential to stimulate useful discussions about how to improve practice. But they should not be taken as direct evidence of the effects of instructional practice.

Notes

[1] Goldstein and Spiegelhalter (1996) discussed this distinction and it emerged in the comments of several of their discussants in an issue of the *Journal of the Royal Statistical Society* that highlighted themes common to those considered in the current issue of the *Journal of Educational and Behavioral Statistics.* Willms and Raudenbush (1989) consider the stability of these effects over time.

[2] We must assume that the number of students so randomized per school is comparatively small lest the influx of new students modify the context, *C,* which is part of the treatment.

[3] A change in *P* could lead to a change in *C* over the long run if, for example, more advantaged parents send their children to a school in order to reap the benefits of improved practice.

[4] One might imagine an experiment in which students are assigned at random to schools that vary on *P* but have the same *C.* While such an experiment would reveal the impact of *P,* conducting it would require that *C* be completely observed. Assigning schools at random to *P* eliminates that strong requirement.

[5] If we assume strong ignorability (that *X* and *W* adequately capture the selection of schools into values of *P*), and that the regression model assumptions hold, then the variance of the estimates of the effect of P_j based on regression is a lower bound on the variance of the Type B effects. The variance of the Type A effect is the upper bound. If these bounds are close together, one can claim to have "bracketed" the variance of the Type B effects. This doesn't help with estimating effects for particular schools, however, and such individual estimates are the object of an accountability system.

References

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., et al. (1966). *Equality of educational opportunity.* Washington DC: National Center for Educational Statistics.

128

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society A, 159*(3), 385–443.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

Lee, V., & Bryk, A. (1989). A multilevel model of the social distribution of educational achievement. *Sociology of Education, 62,* 172–192.

Raudenbush, S., & Willms. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307–335.

Raudenbush, S., Bryk, A., & Ponsiciak. S. (2003, April). *School accountability.* Paper presented at annual meeting of American Educational Research Association, Chicago, Il.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods, Second Edition.* Newbury Park, CA: Sage.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 17,* 41–55.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6,* 34–58.

Willms, J. (1986). Social class segregation and its relationship to student's examination results in Scotland. *American Sociological Review, 51,* 224–241.

Willms, J., & Raudenbush, S. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26*(3), 209–232.

Author

STEPHEN W. RAUDENBUSH is Professor, School of education, Professor, Survey Research Center, and Professor, Department of Statistics and Sociology, University of Michigan, 610 East University, 4109 SEB, Ann Arbor, MI 48109. His areas of specialization are analysis of multilevel data and experimental design.