

The Goals and Uses of Value-Added Models

Derek C. Briggs

University of Colorado at Boulder

November 10, 2008

Paper prepared for a Workshop Held by the Committee on Value-Added Methodology for Instructional Improvement. Program Evaluation and Educational Accountability sponsored by the National Research Council and the National Academy of Education, Washington DC: November 13-14, 2008.

Educational Interventions and Value-Added Residuals

In this brief paper I offer the perspective of someone who has both conducted research on some of the psychometric underpinnings of value-added models (VAMs), and more generally followed the broader research and policy discussions surrounding their usage over the past five years. While the appropriate goals and uses of value-added models in education are subject to much debate and some consensus, it might be worthwhile to begin by framing the current discussion with respect to what a VAM actually does, just from a mechanical point of view. Using the longitudinal test scores of students as inputs, a VAM estimates as an output a numeric residual associated with a specific educational intervention. Broadly defined, an educational intervention represents something that has been implemented with the intent to increase student achievement. I use the term “residual” rather than the term “effect” or “measure” because I think the meaningful interpretation of VAM residuals is equivocal. Some would argue that a VAM residual is an estimate of a causal effect; others would argue that the residual only represents a descriptive measure. I will argue that the interpretation hinges primarily upon two factors: 1) the nature of the underlying educational intervention being parameterized in a VAM and 2) the intended use of the value-added residual subsequently estimated.

Historically most educational research has focused on interventions that are “manipulable” from a policy perspective. By manipulable I mean that it would be relatively easy (though not necessarily cheap) to expose students to more or less of the intervention. Examples would include reductions in class size, the introduction of web-

based learning technologies to a curriculum, and test-based grade retention. Because almost all educational interventions are implemented by teachers and administrators in school settings, for decades VAMs have been used to control for heterogeneity in teacher and school quality through the specification of fixed effects and/or random effects at the teacher and school levels. What has made the value-added methodology simultaneously intriguing and controversial has been a shift in focus since the late 1990s to define teachers and schools themselves as the principle educational interventions of interest.

When a VAM is used to estimate a residual for a more traditionally manipulable educational intervention in a quasi-experimental design context, I think the intended interpretation as a causal effect is relatively straightforward. Indeed the average causal effect of a manipulable intervention has a natural “value-added” interpretation: it is the amount by which a student’s test score outcome differs from what it would have been in the absence of the intervention (i.e., the counterfactual outcome). In contrast, Rubin, Stuart & Zanutto (2004) and Raudenbush (2004) have pointed out that the value-added residuals associated with teachers and schools are very difficult to conceptualize in a causally meaningful way. This is largely because as an educational intervention, the amalgamated characteristics of a teacher and/or school to which a student is assigned are difficult to change—i.e., to manipulate—over a finite period of time. For these and other more technical reasons we are likely to hear from Dan McCaffrey and J.R. Lockwood among other, many researchers would prefer to interpret teacher and school-level VAM residuals as purely descriptive measures, presumably of some aspect of teacher and school quality.

teachers (2) the specific actions taken within school and classroom environments by teachers and administrators to increase student learning. The latter represent interventions that are readily manipulable from a policy perspective; the former are not. The VAM residuals estimated for the former are interpretable as causal effects; the residuals estimated for the latter might be better interpreted as descriptive measures. In each case low or high stakes uses of VAM quantities are possible. When a VAM is used for research purposes that are methodological or exploratory in nature, or as a means of identifying teachers and schools potentially in need of additional resources, there are usually low stakes attached to such interpretations. This will be the case whether the VAM residual is interpreted as a descriptive measure or as a causal effect. On the other hand when a VAM quantity is being estimated for evaluative purposes, the stakes often become quite high. In these cases, I think the distinction between causal effect and descriptive measure might be more important.

When the educational intervention under investigation is parameterized as a teacher or school, the interpretation of the associated VAM residual as a descriptive measure rather than a causal effect shifts the technical conversation from a consideration of *internal validity* to a consideration *construct validity*; from *statistics* to *psychometrics*. That is, if a VAM residual is to be interpreted as a causal effect, the fundamental validity issue from a statistical point of view is whether we can obtain parameter estimates that are unbiased and precise. In contrast, if a VAM quantity is to be interpreted as a descriptive measure, the fundamental validity issue from a psychometric point of view is the extent to which empirical evidence can be provided that collectively supports the intended interpretation and use of the measure. The latter task is just as challenging as

the former, but it is decidedly messier and much less proscriptive as a process. Nor does it guarantee that issues of causal inference can be avoided. At some point, if a descriptive measure is the primary basis being used to reward or sanction teachers, the implied inference that, for example, higher teacher quality produces higher VAM quantities would need to be defended empirically.

For the balance of this paper I will primarily focus attention on the use of VAMs for the purpose of estimating teacher or school-level quantities for either low or high-stakes uses. I pose and briefly address three hypothetical “frequently asked questions” that raise big-picture issues for further discussion. I then offer some concluding thoughts on the prospects of using VAMs as part of a balanced system of educational accountability.

Some Key Questions about Value-Added Modeling

To what extent is the use of value-added modeling consistent with the approach to educational accountability fostered by No Child Left Behind (NCLB)?

The stipulations of NCLB require that all schools receiving Title 1 funds to test their students annually in the subjects of math, English/language arts and science in grades 3 through 8 and at least once during high school. The performance of students within a given school (disaggregated by demographic subgroups) is then evaluated relative to criterion-referenced thresholds for each subject-specific test. Students are subsequently classified into performance levels, e.g., “unsatisfactory”, “proficient”,

“advanced.” By the year 2012, the target is for 100% of students to demonstrate test performance that would place them in the proficient category or higher. To this end, states were asked to specify target school-level percentages of students classified as proficient or higher each year leading to 2012. Each year, if a school’s aggregate percentage is below the target percentage for any student subgroup or test subject, they will have failed to demonstrate “adequate yearly progress” (AYP). High-stakes sanctions are attached to the NCLB law. If a school fails to make AYP in two consecutive years, it must offer parents the opportunity to choose a different public school for their child to attend. After three years of failing to make AYP, supplemental educational services (i.e., tutoring) must be provided for all students eligible for free or reduced lunches. After five years of failing to make AYP, schools become candidates for restructuring by an external agency.

There are two particularly well-known criticisms of the NCLB-based approach to educational accountability. The first is that it is unfair to schools with heterogeneous student demographics. Since socioeconomic status tends to be inversely correlated with test performance, schools with more disadvantaged students will also be the ones with the highest proportions of students classified as unsatisfactory¹. This will be true in a given year even if a school’s students are making tremendous progress. A second criticism of NCLB is that it has resulted in annual criterion-referenced performance targets that are unrealistic. Annual performance targets have been established in a largely arbitrary manner, without the existence of what Bob Linn (2003) has described as an “existence proof.”

¹ This problem is exacerbated by the NCLB focus on subgroup performance.

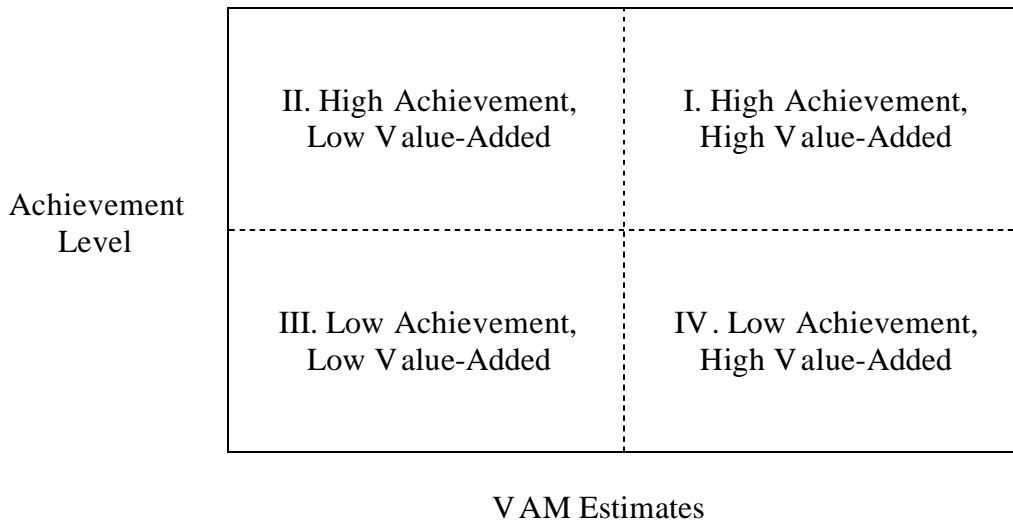


Figure 2. Possible Use of VAM Residuals to Classify Schools in an Accountability System

The incorporation of value-added residuals into the NCLB-based system of accountability would address the first criticism directly by allowing for quantitative distinctions to be made among schools² in terms of two dimensions: achievement level, and achievement growth. This goal is illustrated by the 2 by 2 table in Figure 2. The y-axis in the diagram represents a hypothetical continuum of school-level percentages of students classified as proficient or higher, while the x-axis represents a hypothetical continuum of school-level VAM residuals. The four quadrants are defined by establishing thresholds for, respectively, (1) a satisfactory proportion of students classified as proficient within a given school, and (2) the school-level contribution to growth in student achievement. By contrasting conditional measures of achievement using VAM residuals (x-axis) against unconditional measures of achievement status (y-axis), we can distinguish schools in the 2nd quadrant (high achievement, low growth)

² A parallel arguments could be made here for an accountability system in which teachers were the principal units of analysis. There are however, some important analytical distinctions between the specification and estimation of VAMs in which school residuals are the focus rather than teacher residuals. For details see Briggs & Weeks (2008).

from those in the 1st (high achievement, high growth) and schools in the 4th quadrant (low achievement, high growth) from those in the 3^d (low achievement, low growth).

If such an approach were taken to categorize schools for rewards or sanctions under NCLB, two issues in particular would need careful attention.

1. The reliability of school classifications as a function of value-added estimates. Previous research has found that the precision of school-level quantities is relatively low, which means that in practice only a small proportion can be reliably distinguished as above or below average. Some VAMs produce more precise estimates than others, and this is often a function of data requirements and model complexity. (See next question for more on this.)
2. Reconciling criterion and norm-referenced interpretations of school performance. Recall that the second criticism of NCLB-based accountability focused on unrealistic criterion-referenced expectations of student performance. In contrast, VAM estimates have a normative interpretation as the conditional achievement of one school relative to the conditional achievement of the average school in the system. The latter might be misleading if there are system-wide trends of decreasing or increasing performance³.

For low-stake uses, the two issues above are generally less problematic. Nonetheless it is worth noting that even when no rewards or sanctions are associated with a classification

³ The states of Ohio and Tennessee have begun using a type of VAM based on the work of Bill Sanders that estimates conditional achievement relative to a criterion-referenced standard. The model uses longitudinal student data to project future student achievement, and this is compared to the criterion-referenced performance standards established by NCLB, not to a systemwide average. These differences between projected and expected performance are than aggregated at the school-level.

into “level by progress” quadrants, if such information is made publicly available, schools are likely to take issue with the precision of their classifications when they end up in quadrants II and III.

How Much Data is Necessary? How Complex the Model?

The amount of data that can be used in the specification of a VAM ranges from as little as two years of longitudinal data on a single test subject to five or more years of panel data on multiple test subjects. In general, data requirements depend upon the type of VAM being specified. The Educational Value-Added Assessment System⁴(EV AAS) pioneered by Bill Sanders is probably the most demanding model in widespread use with respect to its data requirements. In theory, the richer the available data, the better the model at controlling for potential sources of bias in estimated value-added residuals, and the greater the ability to produce precise and stable estimates of value-added residuals. When relatively little data is available (i.e., two years), there does not appear to be much difference in the VAM residuals estimated from different types of models (i.e., simple difference score model, fixed effects models, mixed effects models).

One of the positive consequences of NCLB has been the development of a longitudinal infrastructure under which student test performance can be linked to schools (and sometimes teachers) across grades. However, there is a tradeoff in parsimony with attempts to model complex data structures (i.e., data that will often include are large amount of missing tests scores, missing teacher/school links, and transient students) using complex statistical models. As a result it can become very challenging to explain the

⁴ Also known as the “layered” model as described by Sanders, Saxton & Horn (1997).

underlying machinery to educational stakeholders, who are likely to view the model as a “black box.” When stakes are low, it may be more valuable to use less data with a simpler model such that the process becomes more transparent to stakeholders.

Can Currently Available Large-Scale Assessments Adequately Support the Use of VAMs?

As a psychometrician, this is a question I find especially important, and one that is only recently getting sufficient attention. Three key issues:

- Many large-scale assessments were put in place by states very rapidly to comply with the provisions of NCLB. I suspect that most of the grade by grade tests were never designed with the intent to capture longitudinal growth. So when little growth is observed across grades, there are at least two competing interpretations: either the student has learned very little, or test are incapable of capturing it. The less that student test scores appear to grow in any absolute sense, the more difficult that task statistically of disentangling teacher and school contributions to that growth.
- Are vertically scaled tests necessary before a VAM can be implemented? In most cases, this seems to be true, but recently some models have been proposed that allow for value-added interpretations without the need for a vertical score scale (Betebenner, 2008; Mariano, McCaffrey & Lockwood, 2008). Ballou (2008) has recently argued that since vertical scaling practices are unlikely to ever generate score scales with interval properties, it would be most sensible to develop VAMs that only require ordinal outcomes. My view is that the goal of

vertical scaling is first and foremost to make scores comparable in some absolute sense over time. In research conducted with my colleagues Jon Weeks and Ed Wiley, we found that the precision of value-added residuals can be influenced by the way an underlying vertical score scale has been created (Briggs, Weeks & Wiley, 2008).

- One of the most intuitively implausible assumptions of vertical scales that span many grades is that the same unidimensional construct is being measured in each grade. A violation of this assumption is potentially quite important since Lockwood et al. (2007) showed that estimates of teacher value-added residuals are only moderately correlated when they were based on two different subscores from the same large-scale mathematics assessment. This might argue for further research on the possible development of multidimensional vertical scales.

Concluding Thoughts

When debating the goals and uses of VAMs, an important question to keep in mind is “relative to what alternative?” For the purpose of estimating a causal effect for a readily manipulable educational intervention, the alternative to the use of a VAM would be a simple comparison of averages after implementing a randomized controlled experiment. Given a quasi-experimental design, a VAM may be the closest we can come to an approximation of this ideal. On the other hand, in the context of educational accountability, the alternative to the use of VAMs is probably the present NCLB-based

system with its exclusive focus on criterion-referenced changes in achievement levels. As a descriptive measure of school quality, a VAM residual is probably more meaningful than an AYP designation, especially when used in a complementary manner as was illustrated in Figure 2.

What concerns me most is the potential for misuse of VAMs for high stakes purposes. For example, it would be easy for distinctions to be made between schools or teachers on the basis of VAM residuals that are essentially meaningless given the associated standard error. It is also likely that companies marketing a VAM approach will begin making the kinds of claims made by Batelle for Kids, the organization assisting the state of Ohio with its use of VAMs for educational accountability

http://battelleforkids.com/home/value_added/AboutValue-Add :

How Does Value-Added Analysis Improve Teaching and Learning?

Value-added analysis provides important diagnostic information not previously available with traditional achievement reporting. With value-added information...

Teachers are better able to:

- Monitor students' progress ensuring growth opportunities for ALL students
- Predict students' future academic performance
- Modify instruction to address all students' needs
- Align professional development efforts in the areas of greatest need

District and building administrators are better able to:

- Measure the impact of educational practices, classroom curricula, instructional methods and professional development
- Make informed, data-driven decisions about where to focus resources to help students make greater progress and perform at higher levels
- Benchmark progress against other districts and schools
- Identify best practices and implement more effective programs for students

If such claims could be supported, then the use of VAMs would appear to be a genuine educational panacea. However, I know of only one study that has addressed these sorts of claims (McCaffrey & Hamilton, 2007), and the results were at best equivocal. As VAM approaches continue to be implemented in school districts, more validation research of this nature will need to be undertaken. In conclusion, I think there is at least one major consensus with respect to the appropriate use for VAM estimates of school or teacher residuals:

VAM residuals should not be the sole basis for high-stakes sanctions and rewards. They should be used in conjunction with direct observations of teacher and school practices.

References

- Ballou, D. (2008) Test scaling and value-added measurement. Paper presented at the National Conference on Value-Added Modeling, April 22-24, 2008, Madison, WI.
- Betebenner, D. (2008) Norm- and criterion-referenced student growth. Paper presented at the National Council on Measurement in Education Annual Conference, March 26, 2008, New York, NY.
- Briggs, D. C. & Weeks, J. P. (2008) The Persistence of Value-Added School Effects. Paper presented at the 2008 Annual Meeting of the American Educational Research Association, Division D, New York, NY. March 27, 2008.
- Briggs, D. C., Weeks, J. P., & Wiley, E. W. (2008). The Sensitivity of Value-Added Modeling to Vertical Scaling. Paper presented at the National Conference on Value-Added Modeling, April 22-24, 2008, Madison, WI.
- Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007b) The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*. 44(1), 47-68.

Mariano, L., McCaffrey, D., & Lockwood, J. R. (2008) A model for teacher effects from longitudinal data without assuming vertical scaling. Unpublished manuscript.

McCaffrey, D. F. & Hamilton, L. (2007). Value-added Assessment in Practice: Lessons from the Pennsylvania Value-Added Assessment System Pilot Project. RAND Technical Report.

Rubin, D. Stuart, A., & Zannato, E. (2004). A potential outcome view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, 29(1), 103-116.

Raudenbush, S. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29(1), 121-129.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system, a quantitative, outcomes-based approach to educational measurement. In Jason Millman (Ed.). *Grading teachers, grading schools, Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.