

Value-Added Analysis:
Issues in the Economics Literature

Dale Ballou
Vanderbilt University
Oct. 6, 2008

In this paper I discuss issues of concern to economists in the specification and estimation of value-added models. These issues fall under three broad headings: (1) Non-random assignment of students to teachers, and the potential for omitted variable and selection bias; (2) Model misspecification, due to uncertainty about the function describing the relationship between schooling inputs and achievement; and (3) Properties of achievement tests, including the following: how well tests are aligned with the curriculum; ceiling and floor effects; test measurement error and its implications for measured achievement gains; whether test scores are reported on interval scales required for forms of value-added analysis in current use; the timing of test administration, which rarely coincides neatly with the beginning or end of an academic year. I take up these issues in turn.

Non-random assignment

Non-random assignment is pervasive, resulting from decisions by parents and school administrators: residential location decisions, often influenced by the perceived quality of local schools; parental requests for particular teachers, or other efforts to influence teacher assignment; administrative decisions to place particular students with particular teachers, sometimes to improve the quality of the teacher-student match, sometimes as a form of favoritism shown teachers or parents. Because teachers are not assigned a random mix of students, student achievement can be expected to differ from one teacher to the next for reasons unrelated to teacher performance. The challenge for value-added practitioners is to disentangle the influence of teachers from the influence of these other factors.

Value-added analysis typically controls for students' starting levels of achievement, either by measuring achievement in terms of student gains over the period they were assigned to a particular teacher, or by including prior achievement as an explanatory variable in a regression equation. When successful, these methods substantially level the playing field in that teachers are held accountable for students' gains, not for their starting point. However, there is concern that this is not enough, if students differ with respect to their rates of gain for reasons other than the quality of teacher instruction. If so, it is possible that some teachers are systematically assigned "high rate-of-gain" students while others are systematically given "low rate-of-gain" students, biasing value-added estimates in favor of the former.

There is disagreement in the literature concerning the magnitude of rate-of-gain differences and thus the potential bias arising from this source. Researchers that have looked for evidence of a fixed student gain component (evidence that some students are simply fast gainers year in and year out, while others regularly make slower gains) have found that such a component accounts for very little of the overall variance in achievement gains (e.g., Kane and Staiger, 2008). However, this finding may be sensitive to the way achievement tests are scaled (see below). A survey of the literature on education production indicates that most researchers control for differences in underlying rates of gain by including student fixed effects when longitudinal data are available, even when the dependent variable is expressed in terms of gains or when prior achievement is available as a control. The fixed effect in these models removes average differences in rates-of-gain among students before estimating the contribution of other

factors, including teachers. It is therefore thought to ensure a still more level playing field for purposes of value-added analyses.

This belief is too optimistic. First, including student fixed effects in an achievement model produces unbiased estimates of the effects of teachers and other inputs only if the values taken by other regressors in any one period are uncorrelated with the model's error term in all periods. This is known as strict exogeneity (as distinct from zero contemporaneous correlation). In analyses of value-added using data from North Carolina, Rothstein (2007) shows that strict exogeneity fails to hold: past achievement influences future teacher assignments, so that the quality of the teacher a student has in period t is correlated with unobserved determinants of achievement in periods $t-1$.

Second, student fixed effects at best control for only one source of bias—individual students' underlying rates of gain. There are other sources of bias that are more likely to be confounded with teacher effects, notably context variables defined at the classroom level. Peer effects—e.g., the effect of disruptive students on the learning of classmates, or the positive effect of high achieving students on their peers—are one example. Indeed, in most value-added models, any factor influencing achievement that is common to all students of a given teacher will be captured in the "teacher effect." Such effects therefore include spillover from other teachers shared by these students, the impact of administrators and support staff on learning, neighborhood and community effects. For brevity, I will refer to all of these factors as "classroom context variables."

In principle it is possible to distinguish these classroom context variables from teacher quality, if teachers are observed in multiple classrooms. With such data, a common teacher effect can be defined across these observations, allowing for

identification of context effects that vary across classrooms. Progress along these lines has been made by Rockoff (2004) with data on secondary school teachers, each of whom is observed in multiple classrooms within a given year. While this approach is not feasible for elementary teachers with a single self-contained classroom, one can estimate the component of teacher effectiveness that is common across years (say, through the inclusion of a teacher fixed effect in the model) along with coefficients on context variables that change from year to year. Of course, this means the value-added analyst is not estimating changes in teacher performance from year to year, but only the enduring (time invariant) component of teacher effectiveness. If the purpose of value-added analysis is to improve teacher performance, time-varying components of performance are apt to be of great interest, diminishing the usefulness of this approach.

There is still another problem in distinguishing classroom context variables from teacher effectiveness. The covariates available in studies of student achievement are typically only a small subset of the confounding influences on learning. These variables are asked to proxy for a much larger set of family and neighborhood variables in the hope of obtaining unbiased estimates of the effect of school inputs (e.g., teachers). How well they do so depends in part on whether the model also contains teacher fixed effects. Without teacher fixed effects, the zero-order correlation between included covariates and omitted influences determines how well the former proxy for the latter. With teacher fixed effects in the model, what matters is the correlation that remains after removing mean differences in achievement between teachers. This may be much weaker, so that the few covariates available no longer proxy as effectively for omitted variables.

As shown in Ballou (2005), it may be better to estimate teacher effects using a two-stage approach. In the first stage, achievement is modeled in a value-added framework (either as growth, or by controlling for prior achievement). Regressors include all known covariates that influence achievement, but teacher effects are omitted. Partial residuals are then computed from the results, removing the measured influence of the covariates. Teacher effects are estimated in a second stage using the partial residuals as the dependent variable. The result will be biased estimates of teacher effects if teacher quality is correlated with observed covariates. But the bias may be less severe than that which results from weakening the correlation between these covariates and other external influences for which they proxy.

Model Misspecification

While there is no consensus among economists (or other social scientists) about the correct functional relationship between schooling and non-schooling inputs and academic achievement, one influential formulation is that of Todd and Wolpin (2003). Academic achievement of student i in year t is a function of current year inputs plus ability, where the latter is shaped by a student's genetic endowment (typically modeled as a time-invariant student effect) and the cumulative effect of past schooling inputs. The impact of these inputs is assumed to decay over time. This model represents a general framework within which some popular value-added models can be seen as special cases. For example, if there is no decay of the effect of past inputs, then a simple differencing of year t achievement with achievement in year $t-1$ yields a value-added specification in which current year gains are a function of current year inputs, an approach dubbed the "layered" model as one year's inputs (including teachers) are simply layered on top of

previous years inputs to produce the current level of achievement. The Tennessee Value Added Assessment System, since adopted by SAS as the Educational Value Added Assessment System, is a prominent example of a layered model. Although that model is estimated in a levels formulation rather than a gain formulation, the one can be derived from the other. More generally, provided decay occurs at a constant rate across all inputs and time periods, achievement in $t-1$ serves as a sufficient statistic for the entire past history of teacher inputs.

This reformulation is particularly attractive, given that in practice a complete history of past schooling and non-schooling inputs is rarely available to the researcher. However, the assumption that decay occurs at a constant rate across time and across different types of educational inputs is very strong. Moreover, there are some estimation difficulties. First, the model represents the relationship between true achievement and educational inputs. Any given test score is a fallible measure of true achievement, so that inclusion of prior achievement as a regressor raises an errors in variables problem. The conventional solution to this problem is to use instrumental variables that are correlated with true achievement in $t-1$ but not correlated with measurement error. One proposal is to use lagged achievement (test scores from period $t-2$), though this obviously increases the demands on the data. Given high levels of student mobility, in practice many students lack a sufficiently long history of prior achievement to be included in value-added analysis, raising some difficult practical issues. Are only some students to count in value-added analysis? What is the incentive effect if this becomes known to teachers in contexts where value-added is used for high-stakes decisions? Finally, the instruments are invalid if there are omitted (unmeasured) inputs selected in response to past

achievement, a circumstance that seems all too likely if schools, families, or both engage in optimizing behavior.

In order to reduce the potential for omitted variable bias, researchers with longitudinal data often employ specifications that include a variety of fixed effects: e.g., school effects (to control for other schooling inputs); cohort effects (to control for peers). Unfortunately, if teacher quality is correlated with other schooling inputs, including a school effect in the model removes part of the true differences in teacher quality from the estimates of teacher value-added, assigning it instead to the school. This effectively changes the question that value-added analysis attempts to answer. With school fixed effects in the model, teachers are compared only to other teachers at the same school. Each additional effect added to these models further restricts the comparison. When cohort effects are added, teachers are compared to others who taught the same cohorts. The same is true when student fixed effects are added to the model: because mean differences in achievement between students is absorbed in the student fixed effects, groups of teachers with non-overlapping groups of students cannot be compared to one another; indeed, the only information contributing to teacher effects is the variation in achievement over time within each student's history, so that each teacher of that student is effectively compared to other teachers of that student in the same subject. This significantly restricts the inferences that can be drawn about the relative effectiveness of different teachers and may not serve all the purposes for which value-added analyses are intended. For example, if the purpose of value-added analysis is to identify and remove the least effective teachers, models that do not permit comparisons of teachers across different schools are not likely to serve the policy well. For this reason, rather than

school fixed effects, it may be better to use models in which teachers are compared to others who work in schools of a similar type with respect to student demographics, SES, and other indicators of student need.

Finally, virtually all value-added models define the parameter they are trying to estimate (teacher effectiveness) as the increment in the achievement of any given student assigned to this teacher rather than the average teacher in the relevant comparison group, however that group is defined. Thus, the effect of teacher s on student i is assumed to be the same as the effect of teacher s on student j . If, as seems plausible, teachers are not equally effective with all types of students, these models are misspecified. This also raises difficult questions about what value-added analysis should attempt to measure. Is it the effectiveness of a teacher with the types of students to whom that teacher is regularly assigned? Should it be some counterfactual measure of effectiveness, for example, the effectiveness of that teacher with a representative mix of students, or the students to whom that teacher should have been assigned, if administrators made such assignments optimally? Work on these matters is at a rudimentary stage.

Properties of Tests

Of the three topics discussed in this paper, the question of what achievement tests measure and how they measure it is probably the most neglected by economists. Yet teachers' lack of confidence in standardized tests is a major reason they distrust value-added analyses. If tests do not cover what teachers actually teach (a common complaint), the most sophisticated statistical analysis in the world still will not yield good estimates of value added unless it is appropriate to attach zero weight to learning that is not covered by the test.

Some tests avoid ceiling and floor effects, particularly those that are designed to measure student growth across multiple grades; however, others do not. Even if ceiling and floor effects are avoided in the strict sense, tests are not likely to measure achievement gains with equal accuracy at all points on the learning continuum. This is particularly likely in accountability-driven testing regimes, in which the chief purpose of the test is to ascertain whether students have attained a particular level of proficiency. Tests designed for this purpose will attempt to discriminate well among students who are near the proficiency level but in doing so are apt to discriminate poorly among students who are far above or below that level, calling into doubt the suitability of using these tests to measure value added of teachers whose students are in these upper and lower ranges of achievement.

Tests are rarely given at the beginning or end of school years. It is not uncommon for achievement tests to be given in early spring, two to three months before the school year ends. (This is an additional problem exacerbated by accountability-driven testing, inasmuch as states cannot wait until the end of the school year to give tests when the results of those tests are needed by the beginning of the next school year for purposes of accountability.) Learning that occurs beyond the test date shows up, if at all, on next year's test, where, given current practice in value-added analysis, it will be attributed to next year's teacher. Research also shows that students forget at different rates over the summer, depending, among other factors, on family SES. One solution to both these problems might be to test students early in the fall, and use fall scores as the measure of prior achievement for value-added calculations. However, this raises a moral hazard problem if high stakes are attached to gains measured from the fall baseline.

As currently practiced, value-added assessment relies on a strong assumption about the scales used to measure student achievement, namely that these are interval scales, with equal-sized gains at all points on the scale representing the same increment of learning. Many of the metrics in which test results are expressed do not have this property (e.g., percentile ranks, normal curve equivalents). However, this property is claimed for the scales obtained when tests are scored according to Item Response Theory.

This claim requires that examinees and test items constitute, in the terminology of representational measurement theory, a conjoint structure. Unfortunately, it is difficult to confirm that this condition is met. In addition, end users of the data lack access to item-level data to test these assumptions themselves. The best they can do is to check the plausibility of the resulting scales. On this count IRT scales often do poorly. Reasonable rescalings have a substantial impact on students' measured growth.

Practitioners of value-added analysis who are concerned about the impact of test scaling on estimates of value added have sometimes adopted ad hoc modifications of the scale in an effort to reduce scale dependence. These ad hoc approaches do not solve the problem. For example, some methods of test scaling produce scales in which the distribution of ability fans out as grade level advances (as one would expect if there is stable heterogeneity in underlying rates of learning. As we have seen, this kind of heterogeneity can be confounded with teacher and school effectiveness. The Educational Value-Added Assessment System (EV AAS) of the SAS Institute removes this heterogeneity through an ad hoc modification of the scale, wherein reported gain scores are divided by the gain required to keep an examinee at the same percentile of the post-test distribution that he occupied in the pre-test distribution (Ballou, 2007). Thus, if the

distribution in terms of the original scale exhibits increasing variance at higher grade levels, the transformation pulls up gains of examinees whose pretest scores were below the mean and reduces gains of examinees whose pretest scores were above the mean. It is unclear why this should be regarded as superior to using the original scale. It does not solve the problem of scale dependence; rather, it simply substitutes a particular transformation for the original scale. While it is possible to conduct value-added analysis using methods of ordinal data analysis that do not require strong assumptions about test scales, work along these lines is still at a very early stage.

References

Ballou, Dale. 2005. Estimating Teacher Quality from Student Test Scores. Vanderbilt University.

Ballou, Dale. 2008. Test Scaling and Value-Added Measurement. Presented at the National VAM Conference (22-24 April 2008), Madison WI.

March 2004 (revised March 2005) Kane, Thomas J. and Douglas O. Staiger. 2008. Are Teacher-Level Value-Added Estimates Biased? An Experimental Validation of Non-Experimental Estimates. Presented at the National VAM Conference (22-24 April 2008), Madison WI.

Rothstein, Jesse. 2007. Do Value-Added Models Add Value? Tracking, Fixed Effects, and Causal Inference. Presented at the National VAM Conference (22-24 April 2008), Madison WI.

Rockoff, Jonah E. 2004. The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *The American Economic Review*. 94, 247-252.

Todd, Petra E. and Kenneth I. Wolpin. 2003. On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal*, 113, F3-F33.