A Review of Value-Added Models

Terry Hibpshman
Kentucky Education Professional Standards Board

September, 2004

Executive Summary

Value-added methodology (VAM), especially as exemplified by the Tennessee Value-added Assessment System (TVAAS) by William Sanders, has emerged over the past several years as an attractive alternative for evaluating the effectiveness of school systems and school personnel. Its attractiveness stems principally from its purported ability to minimize or altogether obviate a notorious problem in the evaluation of educational data, the complex interactions between student characteristics, community characteristics, school policies, and teacher effects that together contribute to student success or failure. These complex interactions inevitably cast doubt on the results of simpler estimation methods, because it is very difficult to control for all of the relevant variables known to have some effect on student outcomes.

While VAM has been implemented in some places (e.g., the state of Tennessee and the Dallas, Texas school system) as an official teacher evaluation system, and in others (e.g., Philadelphia, Pennsylvania) on an experimental basis, this group of methodologies should not be considered mature or well-formed at this point in its history. No single approach to VAM estimation has proven superior to any other, and there are numerous open questions about the precision of estimates derived from these systems and the appropriateness of rewards or sanctions based on the derived estimates. It can be said with some justice that at this point even the experts on these methods, and of multivariate statistical methods generally, are not in agreement about the optimal model or the appropriateness of VAM as a routine methodology for the evaluation of teacher or school effectiveness.

VAM models, by their nature, are very complex, and a high level of statistical expertise – far beyond that held by a typical administrator at the school, district, or state level – is required to understand them. This lack of intuitive simplicity has contributed to their premature implementation as normative evaluation models, as statisticians with an interest in selling the methodology have glossed over some very real theoretical problems in the interests of simplifying the results so that they can be understood by the consumer. As a result, legislatures, administrators, and other policy-makers often make implementation decisions without an understanding of the limitations of these models.

There are a number of different models in use in this field. Differences in the models stem from efforts by statisticians to resolve the various technical problems that have arisen as the field has developed. None of the models solves all of the known technical problems, and some problems have proven intractable. As a result, while some of the models have proven useful for specific limited purposes, no one can claim to have developed a VAM model of general applicability whose results can be trusted implicitly for the purpose of rank-ordering teachers or schools with enough precision to justify their use in a high-stakes environment.

This report is an effort to identify the major models that have been proposed by various authors in this field, and to assess their applicability for use in Kentucky, should we decide to use VAM to evaluate the performance of teachers, schools or districts. A number of models are described, along with their strengths and limitations, and the implications for their possible use in Kentucky. It is stressed that none of these models are so well-founded that they could be used without great caution to produce any high-stakes results.

Background: Linking assessment, accountability, and teacher performance

Among the numerous innovations that have arisen in the course of the latest round of school reform, perhaps the most visible to the public eye has been the idea of high-stakes accountability (Amrein and Berliner, 2002). Schools, districts, and school personnel have increasingly been subjected to scrutiny, as most states have implemented systems that measure school performance and many have implemented reward and sanction systems. Behind this practice lies the idea that the performance of schools can best be improved by taking regular measurements of the progress of students and assessing consequences for performance.

In Kentucky, as in much of the nation, the accountability system has used a cohort-comparison approach, where schools and districts are held accountable for the average performance of each succeeding group of students. This approach functions by measuring the achievement of groups of students at intervals in their school careers and computing the average of the scores of each cohort as a measure of the performance of the school or district that serves them. While this method has the advantages of relatively low cost and intuitive simplicity, it is not without its limitations (RAND, 2003).[1]

To serve as fair measures of school or district performance, cohort comparisons require the assumption that students in each cohort be similar demographically to students in subsequent cohorts (Sanders and Horn, 1998). If this is not true, then a change in the average achievement of students in a school or district might be due to compositional factors not under the control of teachers or administrators. Aside from the fact that it is unfair to hold individuals responsible for factors over which they have no control, this presents the problem of remedial efforts having a high probability of failing to address whatever problem may be suppressing performance.

This assumption – that succeeding cohorts are similar demographically to the cohorts they follow – is of questionable validity. Populations do change over time, and school districts sometimes reapportion the catchment areas of their schools. There is considerable doubt in any case that we could detect even large-scale shifts in student demographics, given the scarcity of the data available to us about individual students.

Coincident with the recognition of the dependence of common accountability methods on questionable assumptions has been an increasing interest in isolating the specific factors that contribute to educational success. In particular, teacher quality has come to be regarded as a major source of variation in student achievement, and there has been a great deal of interest in the development of methods that can both demonstrate the truth of this belief and measure the effectiveness of particular teachers.[2] This interest has gained considerable momentum in the current high-stakes atmosphere, as interest in the idea of holding teachers individually accountable for the achievement of their students has grown.

[1] For the sake of readability, we will use abbreviations to cite frequently-cited references with multiple authors.

[2] Note that the methods described herein are equally valid for measurement of the performance of schools, districts, teachers, or statewide educational systems. While these methods have been developed primarily in the context of teacher effectiveness research, some studies have used them as measures aggregated at higher levels. Conceptually, students are nested in classes, which are nested in schools, which are nested in districts, which are nested in statewide educational systems, and the analysis can proceed at any of these levels, using the same data.

Thus there has been an increasing interest in the problem of measuring the performance of teachers, schools, and districts independently of factors such as school composition that are related to student achievement, but cannot be easily manipulated. Beginning in the mid-1980's, as a result of advances in statistical methodology, scholars have begun to apply sophisticated mathematical methods to this problem (Raudenhush and Bryk, 1986). The two types of models commonly used are known as *mixed models* and *hierarchical linear models*[3]. Approaches to evaluating teacher effectiveness by the use of these models have come to be known in the context of teacher and school accountability as *value-added methods* (VAM). These methods have undergone considerable development over the past ten years or so, culminating in the implementation of operational high-stakes teacher assessment systems in a number of locations, including the state of Tennessee; the Dallas, Texas and Minneapolis, Minnesota school districts; and in experimental implementations in a number of places, including Philadelphia, Pennsylvania. While a number of different methods have been implemented, the most commonly used, and the one that has received the most attention, has been the mixed-model approach developed by William Sanders, the Tennessee Value Added Assessment System (TVAAS).

Implementation of these systems has not been without controversy. As will be seen in the remainder of this paper, there are numerous open questions about the precision of these methods as measures of teacher effectiveness and the extent to which they control for factors not related to teacher performance. But the controversies surrounding VAM require a great deal of rather esoteric knowledge about multivariate statistics, and are not readily accessible to nonspecialists. As a result, there is increasing pressure on policymakers to implement these models, in the absence of clear information about their limitations.

The purpose of this paper is to answer questions about the potential of VAM as an assessment model in Kentucky, should state policymakers decide to go in that direction. The paper is organized as follows:

In section 1, we look at factors associated with the controversies and unanswered questions surrounding these methods.

In section 2, we look at specific VAM models, their strengths and weaknesses, and potential problems in their use.

In section 3, we evaluate the two previous sections with respect to the question of which model(s) might be appropriate for use in Kentucky and the conditions that would have to exist here for their effective use.

[3] Note that the two methodologies are not mutually exclusive, although they are not identical either. A mixed model might be a hierarchical linear model, and a hierarchical linear model might be a mixed model.

1. Some preliminaries

First it is important to explain some terminology that will be used throughout this paper.

Compositional variables are measures of factors such as socioeconomic status, rurality, average parental education, and the like, that characterize schools and the communities in which they are embedded (TCMLRAFR, 2004).

Policy variables are measures of school practices and organizational rules that operate at the school or district level. These include such things as curriculum alignment, use of technology, school council organization, disciplinary policy, and so forth (Raudenbush and Bryk, 1986).

Fixed effects are measures of factors that are common to schools. (Hilton-Minton, 1995).

Random effects are measures of factors that vary with individuals or schools (Bryk and Raudenbush, 2002).

Bias is the tendency of statistical results when applied to sample data to inaccurately estimate the population value.

Shrunken estimates are estimates of statistical parameters that result from procedures that pull the results toward unbiased estimates of parameter values (Raudenbush and Bryk, 2002). Shrunken effects in general produce more precise estimates than are produced otherwise.

Model specification is the process of identifying the variables of interest in answering a research question, together with a method of analyzing the data (Greene, 2003).

Covariates are variables that are correlated with the outcomes of interest and the explanatory variables, that must be controlled in order to assure that the outcomes have a real relationship to the explanatory variables.

In order to understand the difficulties inherent in the estimation of teacher effects, it is useful to imagine a set of circumstances in which teacher effects would be easy to estimate - the "randomization experiment" proposed by RSZ (2004). What we would want is the ability to collect data in a setting where we had complete control over the sources of variation that usually confound the results of teacher effectiveness studies.

Control of such variables is ideally done by selecting a setting where the fixed effects are the same for all subjects and by assigning subjects at random to treatments. Since the "treatment" in this case would be placement in a particular teacher's classroom, this would mean that we would try to assure that the contextual and policy effects were the same for all teachers' classrooms, and we would randomly assign students to classrooms. If we were interested only in the relative rankings of a particular set of teachers, we would not have to randomly assign teachers to classrooms, but if our interest were more theoretical – if we were perhaps interested in determining whether a proposed methodology could produce precise estimates of teacher effects – we would want to randomly select and assign a group of teachers from the population of teachers of interest.

The design of our randomization experiment would then consist of selecting a single school, which would effectively obviate compositional and policy variables that otherwise would differ from one school to another, and we would assign students randomly to teachers. We might also want to preselect students so as to make them a bit more homogenous than they otherwise might be: we might, for example, select only students with no previously diagnosed disability.

The final step in controlling sources of variance would be to test all students twice, once just before assignment to teachers, and once at the end of the school year. This would assure that there were no systematically varying effects that might occur between the first and second testing occasions, and would assure that the measure of achievement – the difference between pretest and posttest scores – was a relatively pure measure of change brought about by teacher effects.

Having thus controlled for the various confounding variables, estimation of teacher effects would be a straightforward matter: we would just perform an ANOVA, with teacher identities as the treatment variable. A simple F-test for differences between the means of the groups would suffice to determine whether there were differences in teacher effects, the mean of each class would serve well as the measure of teacher effectiveness, and we could easily estimate the strength of the teacher effect (i.e., the proportion of the variance in student test scores explained by differences in teachers) with an intraclass correlation. We could use the standard error of the mean to set confidence boundaries around the mean for each classroom and could thus easily determine how much uncertainty there was in ranks computed by the system. We could also perform various diagnostic tests to determine whether there were anomalous patterns in the data, such as inhomogenous variances between classrooms.

Of course, we lack anything remotely approaching this level of control. Students are not assigned randomly to classrooms; schools vary widely on both compositional and policy variables; and we do well to have even one adequate measure of student achievement per year.

The traditional practice when we lack control of possibly interacting confounding sources of variance is to apply statistical controls via the use of complex multivariate models,[4] and in fact this is what is done by VAM. This is legitimate practice, with a long and honorable history. The problem arises not from the general approach but from specific problems that result from the nature of school organization and practice, and from the statistical methodologies necessary to respond to them.

The first, and perhaps most significant problem, is that of *nesting*. Students are identified only with particular classrooms; classrooms exist only in particular schools; schools exist only in particular districts. Statistical controls ordinarily work because some of the subjects with any given set of characteristics cross the boundaries of groups, but this is not the case in most educational settings. As a result, because students are not randomly assigned to classrooms or schools to communities, the characteristics of students and communities are correlated with classrooms and schools, and our statistical models cannot easily separate compositional effects from teacher effects.[5] It is known,

[4] i.e., by conducting covariance analysis.

[5] Note that this is also the problem with studies that compare the results of public and private schools. There is ample reason to believe that private school students are quite a bit different from public school students, on factors associated with achievement, and it is difficult to disentangle the effects of these confounding variables for public vs. private schools, just as it is for public schools in different communities.

for example, that more successful teachers tend to be able to select their assignments (Goldhaber, 2004; Hibpshman, 2004b), and as a result are more likely to have highly motivated students. It is difficult to determine whether the higher average levels of achievement of their students is due to teacher effectiveness, or to the highly motivated students they teach. The nesting problem, in effect, creates a statistical model where much of the necessary data are missing and where the causes of missing data are systematically related to the matter under investigation (RSZ, 2004).

A second problem has to do with our general lack of information about which variables are important, and deficiencies in available data when we do have such information (RAND, 2003; MLKLH, 2004). We know for certain that student, community, school, and district variables are of importance in determining student achievement, but we are not entirely certain which particular measures are important, and our data systems were not designed to make information about such things readily available. We often use the proportion of students eligible for free and reduced-price lunch as a measure of socioeconomic status, for example, but this measure is at best only a proxy. We usually lack information about such things as parental education and income for individual students, even though these are often viewed as important factors in determining achievement, and are probably better measures of socioeconomic status. Similarly, we know for certain that school policies are important as determinants in educational outcomes (Raudenbush and Bryk, 1986), but we have no widely-accepted theoretical model for how these policies affect learning, nor do we have any data at all about how they have been implemented in schools. As a result, the effect of an important source of variance cannot be controlled by our studies.

A third problem is that of *model specification*. This problem has two aspects. First, an effectively infinite number of statistical models could be used to evaluate school data, and the same data could be processed in any of several different ways. Some models will perform better than others, but because the models are all very complex, simple and clear-cut diagnostic measures are not available, and the performance of these statistical models can only be evaluated via simulation studies. The other aspect of this problem has to do with the missing variables problem noted above: a model that fails to include information about crucial variables is by definition misspecified (MLKLH, 2004; Greene, 2003).

A fourth and final problem is the sheer complexity of the methodology. In the case of simpler statistical methods like our randomization thought experiment above, the results have intuitive appeal: a non-statistician can examine the mean achievement scores for classrooms of students and understand their meaning – the higher the score, the better the teacher. Further, a non-statistician, even if he does not understand the logic of ANOVA, can understand how we arrived at our rankings. Not so when we conduct covariance studies with nested multivariate data: only an individual with a strong background in multivariate statistics can understand how we arrived at our results. This makes it difficult for policymakers to make well-informed decisions about the selection of these methods for routine use.

The various VAM models have arisen in response to the first three problems. As will be seen, these models have not as yet managed to definitively solve these problems, and numerous uncertainties remain concerning the specification of VAM models that would produce high-precision estimates of teacher effectiveness. In the process of attempting to resolve the first three problems, VAM models become increasingly complex, creating the fourth problem. This is the subject of section 2.

2. Value-added models

Most of the information in this section comes from review of a single issue of the *Journal of Educational and Behavioral Statistics* (JEBS), in 2004. This was an issue dedicated to the analysis of the state of the art of VAM, and presents viewpoints of many of the leading scholars in the field. Other reviews of the field exist as well, such as RAND (2003), which will be incorporated as appropriate in the discussion of the views of the various authors of the *Journal of Educational and Behavioral Statistics* articles.[6]

The authors of these journal articles were wrestling with two problems:

1. How to classify the various approaches to VAM that have been suggested by scholars in the field
2. How to incorporate school and student-level covariates into the models

The JEBS articles consist of three major studies that identify and evaluate existing VAM models, and several opinion and analysis articles that react to the major studies. This section will evaluate each of the major studies in turn, and incorporate ideas from the analysis pieces as appropriate. The three studies are TCMLRAFR, BSW, and MLKLH.

Before discussion of these studies, it is important to be clear about exactly what VAM models are intended to do. VAM models attempt to determine how much of the change in student achievement over some time horizon is due to the efforts of teachers, schools, or districts, usually after controlling for student or school factors that differ between classrooms, schools, or districts. The change in student achievement is typically measured from one year to the next, and the outcome variable studied by the models is typically the difference between student achievement at the end of the latest year and achievement at the end of one or more previous years. Most models make some effort to eliminate the effect of a student's achievement history on his or her current year achievement. The results are typically reported as a proportion of gain for a student in a particular classroom, school, or district, compared to students in other units in the analysis. It is important to note that it is very rare for students to show no appreciable gain in an absolute sense: the outcomes are almost always comparisons between gains due to specific units, and even poorly-performing units may demonstrate substantial gains in an absolute sense.[7]

The TCMLRAFR Study

TCMLRAFR (2004) described four different value-added models:

Fixed-effects models (FEM), where school effects (i.e., the improvement in student achievement due to teacher or school efforts) are taken to be fixed rather than random. This is the simplest of all models, requiring little computational complexity and not much mathematical knowledge beyond elementary statistics. This model thus has intuitive appeal to policymakers, since the meaning of the results is much easier to comprehend.

[6] For a review of the RAND study, see Hibpshman, 2004a.

[7] i.e.., we expect every student, even with relatively ineffective teachers, to learn something every year.

An extension of this model, the simple fixed effects model, or SFEM, is an intuitively simple model that incorporates no information about confounding factors, does not apportion variance when students attend multiple schools, and by the nature of the statistics used, does not produce shrunken estimates. This model estimates effect scores for schools by comparing school effects only to the effect sizes for the districts to which they belong.

The layered mixed effects model (LMEM), exemplified by the work of Sanders and Horn, implemented as the TVAAS. It uses the information in non-zero covariance between test scores at different times. This is a model for change scores with random school effects that does not attempt to account for confounding factors at either the school or student level. This method attempts to eliminate the effect of confounding factors by assuming that all such factors are eliminated by the use of multiple measures on each student (BSW, 2004). LMEM models may use measures of student achievement on two occasions, or they may include longitudinal measures over several occasions. LMEM models produce shrunken effects. This model has the added advantage of incorporating information from all available data, even data that has missing values on some variables, and accounts for the effect when students have attended multiple schools. Finally, LMEM allows for the simultaneous assessment of results from multiple content areas.

Hierarchical linear models (HLMM), which assume that school effects are random. These models produce shrunken effects. There are two of them:

- The simple unadjusted change score HLMM (UHLMM) with random intercept. This model does not account for compositional or student-level covariates.
- A demographic and intake score adjusted HLMM (AHLMM), with outcome defined by a change score, and with student and school-level covariates.

TCMLRAFR were interested in determining how much difference there was in the effect scores produced by these various models. They note that considerations that may have importance in theory may make little difference in a practical sense, and if a simpler model produces results comparable to more complex models, it may be preferable because of its intuitive appeal. In order to evaluate the relative appeal of the various models, they performed a simulation study, producing estimates from all the models using a standard dataset.

The results of the simulation study showed very strong correlations (typically > .9) between results generated by SFEM, LMEM, and UHLMM, but much more modest correlation between the results of AHLMM and all other models. TCMLRAFR concluded on the basis of these results that the SFEM performed about as well as the other two models that did not incorporate compositional or student-level covariates, and could be expected to produce similar results at a much lower computational cost. It was noted that these results were based on only two years of student achievement data and that the incorporation of more years of data might affect the relationships among effects generated by the three models.

The difference between AHLMM and all other models was notable, and indicate that when compositional and student-level covariates are included in the analysis, the estimates change. Since it seems likely that such covariates do indeed affect student learning, it is arguable that the AHLMM produces more precise estimates than do the other models. One thus has a choice between using SFEM because of its computational simplicity and intuitive appeal, or AHLMM, which incorporates more information and arguably produces more precise estimates.

TCMLRAFR argue that there are two undesirably polar approaches to school accountability represented by SFEM and AHLMM. SFEM holds schools accountable for student achievement gains regardless of confounding factors, while AHLMM fully excuses schools from responsibility for such factors. In a high-stakes environment, an accountability model based on SFEM might encourage teachers and administrators to migrate to schools and districts with better socioeconomic indices, thus depriving students with the greatest need of the most capable staff. AHLMM, on the other hand, could institutionalize low expectations for poor or minority students.

The BSW Study

BSW (2004) evaluated the TVAAS model (which is equivalent to TCMLRAFR's LMEM model). They noted that studies of the inclusion of contextual factors in HLM models almost always show that the results are sensitive to such effects, and they note that the TVAAS can include context factors if desired. Inclusion of these factors tends to bias measures of school and teacher effects towards zero.

Using data from the vast database accumulated by TVAAS, BSW conducted a simulation study to determine how much teacher effect sizes reported by the TVAAS would change if student and school compositional effects were entered into the model. The simulation study used student eligibility for free and reduced price lunch, race other than white, gender, the two-way interactions between these, and percent free and reduced price lunch by classroom as covariates. Thus, there were three student-level covariates and one school composition variable used in the study.[8]

The conclusion reached by BSW was that student-level covariates showed only a moderate influence on teacher effectiveness scores. The scores produced by the two models were 2.7 times more likely to agree than to disagree in reading, 3.5 times more likely in language arts, and 8.5 times more likely in mathematics.

With respect to the school composition variable (% free and reduced lunch), BSW found that there was a significant effect on the magnitude of teacher effectiveness scores, but they noted that the direction and magnitude of the regression coefficients showed that the relationship between the percent free and reduced variable and teacher effectiveness was unstable, and therefore not much confidence could be placed in the results.

Having concluded that student-level covariates had little effect on teacher performance scores, BSW offered four possible explanations for the result of the simulation study:

1. If the great majority of teachers have roughly the same mix of poor and non-poor students, white and non-white, then adjusting for demographics will not change

[8] The covariates were not selected randomly: BSW selected only those variables from the available alternatives that showed a significant relationship to student outcomes.

estimated teacher effects. This explanation was discounted because we know that the mix of poor and nonwhite children does indeed vary widely from one classroom to another.

2. The impact of student variables is not large enough to make an appreciable difference to estimated teacher effects. This explanation was discounted because if it were true, a similarly modest effect should be found in the fixed effects model as well. Since the results of the fixed effects model are significantly different from those of the TVAAS model, this cannot be true.
3. The high correlation between adjusted and unadjusted effects is caused by shrinkage. This was discounted for technical reasons.
4. Student factors add little information beyond that contained in the covariance of test scores. That is, other test scores contain much of the same information.

The MLKLH Study

MLKLH (2004) were concerned with creating a system for classifying VAM models as a means of specifying the conditions under which one or another would be a valid methodology. Their approach was to specify a “general” model, and then show how different models suggested by themselves or others would be special cases of this general model. The general model incorporates information about the overall school achievement mean, the proportion of schooling provided to students included in the analysis, the proportion of schooling provided by teachers in the model, student-level demographic factors, and school-level compositional factors. The model was then extended to include information about the effects of prior year schools and teachers. It was noted that the first-year estimates for teacher effects are likely to include information about the student’s prior history and should be interpreted with caution, but otherwise this model represented an effort to include every possible effect that might be investigated by a VAM model. This model is most similar to the AHLMM model described by TCMLRAFR.

Having described this general model, MLKLH then described four special cases:

Covariate adjustment models, similar to models used in the economics production literature. These models use previous scores as covariates for current outcomes. Student and compositional variables could be included in this model. They produce biased estimates when the covariate and residual error terms are correlated.

Repeated cross-section models of gains. This type of model assumes that all scores are on the same scale. It uses difference scores from adjacent grades to produce gain scores. Like the covariate adjustment model, student and compositional effects can be accounted-for.

Cross-classified models. These explicitly model the cross-grade correlations and the effects of multiple years of teachers on student outcomes. Student growth over grades is modeled as a linear trend. They explicitly model the cross-grade correlations and the effects of multiple years of teachers.

The TVAAS layered model

Most of the analysis of the possible uses of these various models is quite technical, beyond the scope of the present paper. There are, however, a few issues that are of central importance to our task:

1. MLKLH note that omitted variables that are randomly distributed should have little effect on the results of any of the models. But when omitted variables cluster by class, or when they differ by strata, neither the general model nor its special cases is capable of disentangling teacher effects from the effects of student-level covariates.
2. The Cross-classified and layered models are most sensitive to the effect of omitted variables.
3. The Cross-classified and layered models use data from all students, even those with incomplete records. This is an advantage of these models, but precision is reduced and bias is introduced when data are not missing completely at random (MCAR).
4. The TVAAS makes the assumption that teacher effects persist uniformly into the future, but this may not be true.

To investigate the effects of the various technical issues related to the models, MLKLH performed a simulation study that compared the results of the general model with those of a layered model similar to the TVAAS. They found significant differences between the estimates of the persistence of teacher effects and significant differences between the estimated teacher effects for the two models. Most significantly, the size of the teacher effect varied with grade for both models, but the general model suggested that teachers contribute less as students progress through schools, while the layered model suggested the opposite. The general model was found to have a significantly better fit to the data than the layered model.

3. What is the best model, and how should we use it in Kentucky?

In constructing the two sections above, I was careful to avoid as much as possible a discussion of the technical mathematical features of the various models, as described by the authors of the cited studies. While this was necessary in the interest of making the material accessible to nonstatisticians, it had the disadvantage of making it difficult to capture the full flavor of the discussion by the authors and required that I altogether ignore some of the features of the models that make them more or less attractive for particular purposes. For this reason, I begin this section with some general comments about the field of VAM and the different models, which are offered without proof. I invite readers to consult the articles themselves for additional details.

The first feature of all of these models that emerges from my review is that there is little reason to place confidence in the precision of estimates that derive from them, in a high-stakes sense. This was a conclusion drawn by RAND (2003), and a reading of the studies makes this abundantly clear. This lack of confidence derives from the following:

1. Model specification is always in doubt. At present, the authors in this field can describe in general what would happen if particular aspects of model specification were violated, but cannot easily determine when a model has been misspecified.
2. There are numerous possible sources of bias for any of these models, and they are difficult to diagnose. In any case, the experts in the field are not in agreement either about the various sources of bias or their severity.
3. Most if not all of the models, in order to avoid the problems leading to bias, must make what one reviewer called "heroic assumptions" (RSZ, 2004). If these assumptions are not valid, then the models are generally misspecified.

This is not to say that these models are entirely without value as measures of teacher or school effectiveness. The authors of the RAND study concluded that VAM methods were capable of accurately identifying teachers at the extremes of the performance distribution but were not able to rank-order teachers in the middle of the distribution with any precision. There are indeed valuable uses of VAM models, especially as research tools, if we keep this in mind.

The problem of evaluating the effect of student and school-level covariates on teacher effectiveness is particularly daunting. I am inclined to see the arguments of MLKLH as more compelling than those of BSW. BSW claim that most of the variance due to these factors is accounted for by covariance of test scores (i.e., as Sanders and others suggest in other places, each student serves as his own control), and that the effect of including student and school compositional variables on teacher effectiveness models is relatively small. MLKLH however make a compelling argument for the idea that separation of these effects is an intractable problem, given the state of the art. In addition, I find less than compelling the BSW argument that the effects of modeling covariates, as demonstrated by their own simulation model, are "moderate". They found for reading that models with and without covariates were 2.7 times more likely to agree than to disagree, 3.5 times more likely for language arts, 8.5 times more likely for mathematics. If expressed as percentages, these translate to 27%, 22%, and 10.5% disagreement, respectively. While these proportions do show substantial agreement among the models (73, 78, and 89.5%), one is led to wonder whether a misclassification rate as high as 27% (or even 11%) is an acceptable level of precision in a high stakes environment. Moderateness exists in the eye of the beholder.

In any case, given what we know about the relationship of demographic characteristics of persons to their educational attainment, it seems unreasonable to think that covariates would have no relationship at all to outcomes independent of teacher effects. BSW suggest four possible explanations, three of which they discount, although they do acknowledge that this is not necessarily an exhaustive list. Keeping in mind the MLKLH finding that it may be impossible to fully separate teacher effects from student and school covariates, I would add two possibilities to their list:

1. We are simply including the wrong covariates in our models.
2. We lack adequate data to include additional covariates, even though we know or suspect they are important.

This latter possibility was suggested by more than one of the authors reviewed for this paper (see e.g., Raudenbush and Bryk, 1986 on the unavailability of school policy data), and was clearly a problem with which all of the authors struggled. The truth is that the data used in these studies are often those that are available, rather than the data we would like to have, and one is led to suspect that the often equivocal results might be due to nothing more than that.

It is important to note the MLKLH criticism of BSW's claims about the persistence of teacher effects. Sanders and others who work with him make rather extravagant claims about the persistence of teacher effects over several years. As MLKLH note, the high rates of persistence reported by BSW and others who use similar methodologies result from assumptions that may not be supportable by the existing data. When MLKLH modeled this problem using less stringent assumptions, they found much more modest persistence effects. This casts doubt on the claims of those who favor the TVAAS and related models, and suggests that the effects of poor teaching may be more remediable than has been previously believed.

In sum, after considering all of the various models suggested by the authors of the cited references, the following conclusions seem reasonable:

A. None of the models can be shown to have sufficient precision to be useful in a high-stakes environment.[9]

B. Models that do not incorporate student or school-level covariates produce estimates that can reliably identify teachers, schools, or districts at the extremes of the effectiveness distribution, but will be biased in favor of teachers who work with more-advantaged populations. Using multiple years of

[9] This raises what is likely to be a concern in the future. Individuals subject to these methods in a high-stakes environment could well initiate legal action over their lack of precision, and equity issues related to the fact that VAM methods are always applied to just a subset of teachers. A search produced no existing case law on the subject, but this field is in its infancy, and as VAM models proliferate, it is probable that lawsuits will as well.

data as in the TVAAS layered model will only partially control for this problem.

C. Hierarchical linear models that incorporate covariates (the AHLMM model of TCMLRAFR, or the general model of MLKLH) will produce estimates of teacher or school effects that are arguably more precise and less biased, but will not altogether eliminate the problem. These models are much more demanding mathematically and require significantly more computational resources.[10]

D. The more complex the model, the more convergence is likely to be a problem[11]; the simpler the model, the more specification bias is likely to be a problem, and the more biased estimates are likely to be in general.

E. The problems of inefficiency and bias may be less serious in practice than in theory, and if covariates are not included in the model, a simple model such as SFEM may be effectively equivalent to a more complex model such as LMEM.

This last point suggests a strategy for determining which model is most attractive: if we only want general estimates of the relative effectiveness of teachers and schools, we would do well to use SFEM, because of its simplicity and intuitive appeal; if we are concerned about controlling for covariates, we should use AHLMM, which is functionally equivalent to the MLKLH general model. In no case should we expect to be able to assess rewards and sanctions using any of the methods, except possibly for scores at the extremes.

Carter (2004), in his rejoinder to the analysis by other authors, suggested that there are two policy extremes that can be pursued by accountability systems. On the one hand, using SFEM or some other model that does not account for covariates, we could hold teachers and schools accountable for student achievement, regardless of the characteristics of the students they serve.[12] On the other hand, we can apply a model like AHLMM, which holds teachers and schools accountable only for the amount of student achievement under their control when other factors have been accounted for. The former policy alternative might well have the effect of making it difficult to recruit good teachers to schools with less capable students; the latter could have the effect of perpetuating low expectations for students at the low end of the socioeconomic ladder.

[10] Although, as the authors note, there are many fine packages that can perform these studies, some of them freeware, and this is much less of a concern than it would have been when these studies required mainframe resources. The problem here is really more of having someone on staff with the time and expertise to work with these very complex models.

[11] Convergence is a technical feature of the mathematical methods used by these models. Convergence amounts to the process of beginning with an approximate solution and progressing to a more precise estimate. There is never any guarantee that a numerical mathematical problem will be well behaved in this sense, and the more complex the model, the less likely it will be well behaved. At the extremes, this may mean that no reliable solution is possible.

[12] While it is not a value-added system, this is the philosophy of the Kentucky accountability system.

Kentucky Implications

It is important to acknowledge at the outset that implementation of even the simplest of these models in Kentucky would require substantial changes in our accountability system. Currently, we do not collect data for consecutive years on any measure of student achievement, and we lack the ability to easily track students from one year to the next (Hibpshman, 2004a). All of the models require that both be true. Correction of this problem would require that we collect a yearly measure of some content area for every public school student in some grade range. If we were to opt for the type of system implemented elsewhere, we would limit VAM implementation to a subset of grades and content areas - for example only self-contained elementary teachers and only in mathematics and reading. If we to desire to evaluate teachers at other levels, and in the full range of content taught in Kentucky schools, then the accountability system would have to be altered to measure the achievement of students every year in many different content areas, and which content were measured would depend on which subjects were taken by particular students. Aside from the probable high cost of such a scheme, this would impose an additional nesting condition, since after middle school, students' course-taking depends substantially on student characteristics.

If we were to alter the accountability system to solve these problems, then the question of which model to use becomes a matter of policy. If Kentucky decides to continue with the current philosophy, holding schools accountable for student achievement in an absolute sense, then we should implement something like the SFEM of TCMLRAFR; if we decide that it would be more fair to adjust for covariates, then we should use something like their AHLMM.

Should we choose to use covariates to adjust for school and district fixed effects, or student random effects, then we need to give careful consideration to which specific effects should be included in our models. As noted above, one of the more serious difficulties with VAM in general is that adequate measures of these variables are generally lacking, and this is as true in Kentucky as elsewhere. Further, while we do have some information, as represented in the student-level CATS results, we have no reason at this point to believe that the information collected is the information we would need. We are fortunate in Kentucky in that we have a statewide uniform data collection system at the school level (STI), which can in principle guarantee that the same information is collected on all students, but considerable work remains to be done before we can link this rich source of information about school and student covariates with any possible accountability system. In order to implement something like AHLMM, we would have to solve a number of technical problems in data collection and transmission, and we would have to make a thorough analysis of the quality and completeness of the information available via the STI system.

We should not attempt to use these models to apply rewards and sanctions on a broad scale to any large number of teachers or schools. We could use them productively to identify teachers and schools that are especially effective or egregiously ineffective. If VAM measures were used in this way, it would be advisable to use them in conjunction with other measures of school or teacher performance.

References

Amrein, A.L. and Berliner, D.C. (2002) High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, 10(18). http://epaa.asu.edu/epaa/v10n18

Ballou, D. (2004) Rejoinder. *Journal of Educational and Behavioral Statistics*, 29(1) 2004, 131 – 134.

Ballou, D., Sanders, W., Wright, P. (BSW) (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29(1) 2004, 37 – 65.

Bryk, A.S. and Raudenbush, S.W. (2002). *Hierarchical Linear Models: Applications and data analysis methods*. 2002, Sage Publications.

Carter, R.L. Rejoinder. *Journal of Educational and Behavioral Statistics*, 29(1) 2004, 135 – 137.

Crane, J. (2002). The promise of value-added testing. 2002, Progressive Policy Institute.

Drury, D. and Doran, H. (2003). The value of value-added analysis. *Policy Research Brief*, 3(1), 2003. (National School Boards Association)

Goldhaber, D. and Anthony, E. (2004). *Can Teacher Quality Be Effectively Assessed?* 2004, University of Washington

Greene, W.H. (2003) *Econometric Analysis, Fifth Edition*. 2003, Prentice-Hall.

Hibpshman, T.L. (2004a). Review of *Evaluating Value-Added models for Teacher Accountability.* Kentucky Education Professional Standards Board.

Hibpshman, T.L. (2004b). A Review of Goldhaber and Anthony, *Can Teacher Quality Be Effectively Assessed?* Kentucky Education Professional Standards Board.

Hilden-Minton, J.A. (1995). Multilevel diagnostics for mixed and hierarchical linear models. (Dissertation) 1995, University of California at Los Angheles.

McCaffrey, D., Lockwood, J.R., Koretz, D.M. and Hamilton, L.S. (RAND) (2003) Evaluating Value-Added models for Teacher Accountability. 2003, Rand Corporation

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A., and Hamilton, L. (MLKLH) (2004) Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1) 2004, 67 – 101.

Mislevy, R.J. (1996). Evidence and inference in educational assessment. CSE technical report 414. 1996, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

Reckase, M.D. (2004) The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics*, 29(1) 2004, 117 – 120.

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

Raudenbush, S. W. and Bryk, A. S. (1986) A hierarchical model for studying school effects. *Sociology of Education*, 59, 1-17.

Rubin, D.B., Stuart, E.A., and Zanutto, E.L. (RSZ) (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, 29(1) 2004, 103 – 116.

Sanders, W.L. and Horn, S.P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, 12(3), 1998, 247-256.

Tekwe, C.D., Carter, R.L., Ma, C., Algina, J., Lucas, M.E., Roth, J., Ariet, M., Fisher, T., and Resnick, M.B. (TCMLRAFR) (2004) An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, 29(1) 2004, 11 – 35.