Link to original paper- http://www.rand.org/publications/MG/MG158/

Evaluating Value-Added models for Teacher Accountability

By Daniel McCaffrey, J.R. Lockwood, Daniel M. Koretz, and Laura S. Hamilton
Rand Corporation

In a monograph published in 2003, McCaffrey, Lockwood, Koretz, and Hamilton of the Rand Corporation review the available literature on the subject of Value Added Methodology (VAM) as a measure of teacher effectiveness. This is a comprehensive and painstaking review of the subject, and should be read in its entirety by anyone interested in this group of methodologies. The purpose of the present paper is not to recapitulate the results found there, but instead to evaluate how the results of the Rand study should be used by researchers and policymakers in Kentucky as we think about how Value Added methodology might be used to address questions related to teacher effectiveness in the Commonwealth. To this end, we will answer the following questions:

1. What are the characteristics of Value-Added methods that make them attractive to researchers and policymakers?
2. What is the state of the art of Value-Added methodologies? Is the technology sufficiently developed to justify its use in routine practice?
3. Could a Value-Added methodology be used productively in Kentucky?

1. What are the characteristics of Value-Added methods that make them attractive to researchers and policymakers?

It seems incongruous that we would imagine that teachers have no effect on the learning of their students. It is obvious that without teachers, less would be learned in school than would otherwise be the case, and in fact we employ teachers because we believe this to be true. It also seems obvious that some teachers, for reasons related to training, natural talent, or other factors, will be more effective (i.e., will cause their students to learn more) than will others. It is this latter belief that engenders the interest in Value-Added methods. Given the widespread public discussion over the past twenty or more years about the quality of the nation's teacher force, it is of interest to us to determine which teachers are more effective than others, so that we can identify what factors contribute to effectiveness, and hopefully mold a teacher workforce that exhibits those factors.

The problem historically has been that it is difficult to separate the effects of teachers from the effects of other factors known to be important in student achievement. Characteristics of students and communities such as parental income and education, ethnicity, rurality, and the like are known to have a strong effect on student achievement, and because students are not randomly assigned to either schools or teachers, it is difficult to separate the effects of such factors from those of teachers.[1] Given the interest

[1] For example, consider that students usually attend schools close to the communities in which they live, and as a result achievement will be heavily influenced by community factors, which will be associated with most or all of the students of teachers who work in a particular school. Comparisons made across students from communities that are different on socioeconomic factors, in the absence of controls for such factors,

over the past few decades in the idea of achievement test-based accountability, this presents us with a serious problem: we cannot hold an individual teacher, or perhaps even a school or district, accountable if we cannot show that student achievement is within their control. Without some method of evaluating the effect of individual teachers, we can hold no particular person accountable for educational outcomes.

But recent reform efforts such as the Kentucky Education Reform Act (KERA) make a critical assumption that requires that teachers and schools be held to a high level of accountability: the assumption that all children can learn at high levels. While the definition of "high levels" is rather vague, the structure of the Kentucky accountability index system, as well as the interest of the Kentucky Department of Education in bringing all districts and schools to "proficiency"[2] implies that there is a minimum level of academic achievement that is desirable, which should be attainable by any school or district, regardless of the characteristics of its students or the communities in which they live. These assumptions of KERA and of the accountability system presuppose that districts and schools can make changes to their curricula, teacher skills, or teacher assignments that can be effective in improving student performance, but in the absence of information about which specific teacher characteristics are effective, leave public school authorities in a bit of a quandary about how to best optimize their accountability scores.

Value-Added methods were developed to deal with this quandary. There are really two threads in this field, a thread exemplified by Sanders, which is focused on the assessment of the effectiveness of particular teachers, and a more theoretical thread exemplified by Hanushek that focuses on the evaluation of teacher characteristics that contribute to student achievement. The former has gained the greatest interest, because of its apparent utility in staff assessment[3], while the latter has been of more interest to researchers. The two approaches use different models and methods, but neither uses a simple methodology whose logic is readily accessible to the nonspecialist.

In brief, then, VAM methods are of interest to researchers and policymakers because they hold out the promise of improving the performance of educational systems by identifying those characteristics of teachers that contribute to success, and by identifying teachers that have those characteristics.

2. What is the state of the art of Value-Added methodologies? Is the technology sufficiently developed to justify its use in routine practice?

The short answer to this question is, the claims of developers of Value-Added methods notwithstanding, VAM methods as currently developed are of limited usefulness as a tool

---

will almost always make teachers from less affluent communities look like they are performing less effectively than teachers from more affluent communities. Similarly, it is usually true that teachers within the same school often do not have randomly assigned students, and as a result such comparisons, in the absence of statistical controls, cannot even be made within schools.

[2] i.e., the point at which the average student in each school and district is proficient, as defined by the accountability system, in the academic subjects specified in the Core Content for Assessment.

[3] i.e., because if valid it provides educational administrators with a tool for managing the performance of their employees.

for any routine assessment purpose, but are well enough developed to be quite useful as research tools. In particular, it seems apparent that currently available VAM methods should not be used for high-stakes assessment purposes. [4]

This conclusion is inescapable given the nature of the statistical methods used by VAM methods, and limitations of the data used in these studies. The specific issues involved are quite technical, but can be summarized as follows:

a. It cannot be demonstrated in the case of any VAM model currently in use that all relevant information has been considered by the model.

As noted above, the impetus behind the development of VAM methodology was an effort to solve the problem of controlling for student, school, community, and other non-teacher variables that affect learning but are not within the immediate control of teachers. There is a large number of known factors in these categories, including economic factors associated with particular communities, urbanicity/rurality, student ethnicity, parental characteristics (income, education, etc.), school leadership and stability, and much more. As demonstrated convincingly by McCaffrey et al., the way these factors are used in VAM studies affects the validity of conclusions drawn from them.

There are two problems with the control of these factors in VAM studies. First, no VAM study can, given the present state of knowledge in this field, ever demonstrate that it has included information about every factor that may be of interest. Because this is true, it cannot be shown how much of the results are due to teacher characteristics, and how much might be due to factors that simply weren't evaluated by the model. As a result, student achievement can never be shown conclusively to be due to individual teacher effectiveness.

The second problem is caused by limitations in the data upon which VAM studies are based. Very often, these data sets were not originally developed for this purpose, and may not include information about factors that are known to be important in student outcomes. Most states now collect information on individual students, but only a limited number of variables are ever collected, and the data systems that collect them were designed more for demographic reporting than for teacher evaluation. Aside from this design problem, there is also the fact that some information, such as parental education or income, may not be easily collected, or if collected may be seen as an overly intrusive effort by the state and public school districts.[5]

---

[4] In the words of McCaffrey at al. (p. xx), "The research base is currently insufficient to support the use of VAM for high-stakes decisions … However, it is not clear that VAM estimates would be more harmful than the alternative methods currently used for test-based accountability."

[5] Existing VAM systems attempt to deal with this problem by using repeated measures of student achievement, so that "each student serves as his own control", but there is doubt whether this approach is fully effective in controlling for non-teacher factors.

b. Because of limitations of the statistical methods currently used by VAM models, there is too much uncertainty about the rank-ordering of individual teachers to permit high-stakes decisions to be made on this basis.

It should be noted that McCaffrey et al. conclude unequivocally that VAM methods are useful, and validly demonstrate that there are differences in effectiveness among teachers, and that these effects persist over time. The difficulty arises when one tries to rank one teacher vs. another, or to compute an effectiveness measure for any given teacher. Their point is that these methods are not well enough developed yet to permit these kinds of fine distinctions to be made. They note that the rather extravagant claims made by some of the authors of these studies are not well supported. Because of uncertainty due to the complexity of the statistical models, it is not at all clear how much of the variance in student outcomes is accounted-for by teacher effectiveness: the number might range from a few percentile points after assignment to a series of effective teachers, to as many as 50 points or more.

In addition, they note that there is enough uncertainty in the statistical models to prevent us from rank-ordering teachers with enough certainty to justify making reward and sanction decisions. They do note that the methods are probably good enough to assure that a teacher who is identified as being in the most extreme ends of the distribution (say, in the top quartile or the bottom quartile) probably is either a high or low performer. For teachers in the middle of the distribution, however, there is quite a bit of uncertainty about the exact rank ordering of effectiveness.

These problems are a serious impediment to the use of VAM for high-stakes assessment purposes, but are a much less serious problem for their use as research tools. As McCaffrey et al. note, VAM studies have demonstrated that teacher effects can be isolated from the effects of other factors, and because this is true, these methods, despite their limitations, can be used effectively to identify teacher characteristics that are related to improvements in teacher performance. In fact, several recent studies (like Goldhaber's 2004 paper on NBPTS certification) use this methodology for this purpose, to excellent effect. One can argue that VAM methodology represents a remarkable improvement in the field of teacher effectiveness research, that makes it possible to answer questions that heretofore have not been readily subject to analysis.

3. Could a Value-Added methodology be used productively in Kentucky?

This question is meant to address the possible use of VAM in a high-stakes rewards and sanctions sense, rather than as a research tool. It is also meant only to address the use of VAM on a statewide basis, rather than its possible use by local public schools or districts. Framed in this sense, then, it is a question of whether, within the current assessment methodology used in Kentucky, or some contemplated alternative, VAM methodology could be used productively in the near future as a measure of the performance of individual districts, schools, and teachers.

The answer to this question, given the results of the McCaffrey et al. paper[6], is that it would be hard to contemplate a set of circumstances given the current assessment system in force in Kentucky, where VAM could be used productively; and substantial changes would have to be made in the CATS system before this could happen. The Kentucky assessment system is designed to evaluate the performance of student cohorts within schools and districts, and is not suited for use in teacher-evaluation studies of any kind.

The reason for these results is that CATS does not provide for year-to-year comparisons of student academic achievement. CATS assessments are intended to assess the performance of students at three points in their K-12 school careers, and to that end measurements are taken roughly four years apart. A student will of course have had more than one teacher during any four year time period, and consequently the effects of individual teachers cannot be disentangled from these single-point-in-time measures.

Additionally, because Kentucky does not have a statewide student identifier, it is difficult to track the achievement of individuals over time, even within the limits of the four-year assessment cycle. Consequently, we have, for most students, only a single point-in-time measure of achievement. This may be sufficient to determine whether schools are making progress in the aggregate, but is not granular enough to determine whether particular teachers are more effective than others.

There are VAM models that permit inferences to be made on the basis of a single point in time measure, but none of the existing teacher-evaluation models fall into this category. All existing such models require at least two, and preferably three years of achievement data on individual students, and in the absence of this data pattern are subject to severe methodological criticism.

The magnitude of changes in Kentucky's assessment system necessary to make VAM assessment a viable option would require implementation of yearly achievement testing in all subjects. All subjects would have to be tested, unless we want to evaluate the performance of only teachers in particular subjects, an equity nightmare in a high-stakes system. Given the very high cost of even the quadrennial CATS testing system, it is doubtful that the state could easily opt for this more frequent time horizon.

Summary:

Value Added methods represent a significant improvement in teacher effectiveness research methodology, and are already producing important advances in our understanding of the relationship between teacher characteristics and student achievement. The statistical methods are quite complex, and are not accessible to nonspecialists. The use of Value Added methods for the purpose of evaluating the effectiveness of individual teachers is more problematic. The state of the art has not advanced to the point where such evaluations are likely to be precise enough to meet reasonable requirements for equity and precision, except in cases where we are interested only in teachers at the extremes of the distribution of effectiveness. These limitations

[6] Aside from the general results in section 2.

notwithstanding, VAM methods could not be effectively used on a statewide basis in Kentucky because the existing accountability system is not engineered to provide the data necessary to conduct VAM studies.