INDIANA UNIVERSITY

SCHOOL OF EDUCATION
Bloomington

Using Value Added Models to Evaluate Teacher Preparation Programs

White Paper
Prepared by the Value-Added Task Force at the Request of University Dean Gerardo Gonzalez

November 2011

Task Force Members: Y. Chiang, C. Cole, G. Delandshere, R. Kunzman, C. Guarino, D. Rutkowski, L. Rutkowski, D. Svetina, X. Yuan & Y. Zhou

Overview and Policy Goals

State and federal governments are beginning to introduce legislation mandating the use of value-added models to evaluate the quality of teacher education programs based on changes in student achievement scores across time. One such state is Indiana, where the state's federal application for a waiver from the provisions of No Child Left Behind (NCLB) indicates that student growth data will be used to evaluate teacher preparation programs in a manner modeled after a program implemented in Louisiana. As one of the major producers of teachers in the state Indiana University is an important stakeholder in this enterprise. Therefore, this white paper is written to assist stakeholders to understand the conceptual issues that will need to be considered when interpreting the results that these models yield.

In addition to its intentions to use student growth data to evaluate teacher preparation programs in Indiana, the state's NCLB waiver application also indicates that in collaboration with institutions of higher education its evaluation framework will be taught in teacher and principal preparation programs. As such, the state should adopt a policy explicitly calling for close collaboration with state institutions of higher education and specifically with schools of education to conceptualize and design a teacher preparation evaluation system that can accommodate differences in the programs and the schools they serve. Close collaboration is needed in order to ensure the credibility and effectiveness of the evaluation system to be built for Indiana and taught in the university curriculum. Not unlike Indiana law that now requires school corporations to develop a system of evaluation for individual teachers using multiple measures, a teacher preparation evaluation system should take into consideration the multiple purposes of schooling and education and not be limited to an exclusive focus on test scores. Consistency with the teacher evaluation systems currently being developed by school corporations is important. Finally, policy makers should aim at transparency and make all aspects of teacher preparation evaluation explicit and all data available for peer-review, reanalysis and further study. This white paper addresses the general challenges associated with the use of value-added models and specifically considers the Louisiana model that the state has indicated will be the basis for a teacher education evaluation system in Indiana. Recommendations are made for taking advantage of what is known about such models in order to build the best possible teacher preparation evaluation system for the state.

History and Context

Value-Added Models (VAMs) are complex statistical models, originally developed by William Sanders in the context of Agricultural Genetics at the University of Tennessee[1] in the 1970's. The origin of VAMs is an important piece of information for understanding the logic of these models, which can be explained briefly as follows. Given a sample of land plots that have similar characteristics (e.g., soil quality, sun exposure, precipitation), researchers can assign them randomly to crops, fertilizers, irrigation systems and so on, to study the effect of these on growth. The logic of experimentation rests on the assumption that *everything else being equal* (plot characteristics), the difference in treatment (e.g., different fertilizers) will explain the differences in growth. That is, the difference in treatment can presumably be regarded as the *cause* of the difference in growth.

Beginning in the early 1980's Sanders tried to convince government officials that his model could be used to evaluate teacher effectiveness based on the increase or growth in student achievement test scores from year to year. He was eventually successful and in 1992 Tennessee[2] mandated that Sanders' model be used for all school districts in the state with the aim of developing a more equitable funding system for the schools. The Tennessee Value-Added Assessment System (TVAAS) is still in use today.

Current Uses of Value Added Models

Currently, a few states and many school districts across the US are using or considering VAMs to evaluate teacher quality and/or primary and secondary schools based on changes in student standardized test scores. The federal government has emphasized the importance of measuring teacher performance based on student achievement growth in its Race to the Top competition and has recently announced the provision of funding for states to develop new accountability measures for teacher preparation programs[3]. Currently, Louisiana appears to be the only state intending to implement a statewide VAM to evaluate teacher education programs, starting in 2012-2013. A few states have applied for Race to the Top funding with the intent to develop a VAM for the evaluation of teacher education programs and yet in other states collaborative agreements between school districts and teacher education program providers (e.g., California State University System, New York City, Florida, Denver Public Schools) have been established to improve the support and quality of candidates in these programs. Researchers have also developed VAMs for the purpose of studying these statistical methods, their technical quality and the validity of the claims made based on these models.

Evaluation of VAMs

Most research on VAMs has been conducted in the context of models used for the evaluation of individual teachers' effectiveness based on the changes in their students' achievement scores. Many concerns have been raised about the adequacy of using these models to draw conclusions about individual teacher quality – most identified issues point to a breakdown of the logic of experimentation on which Sanders and others have developed VAMs and the assumptions on which these are based. The following is a partial list of problematic issues and assumptions:

1. Students are not randomly assigned to teachers or to schools and teachers are not randomly assigned to classes or to schools (like plots of land to fertilizers). Therefore, the condition *everything else being equal* cannot be met and causal inferences (e.g., teachers are the cause of student learning) cannot be drawn.
2. Student learning is affected by many other factors than the teacher – (e.g., school resources, curriculum, school climate, poverty, health, dispositions, interests, motivation, prior learning experiences, home environment, community support, peer group) – which are difficult to measure and to control for statistically. As such, VAMs generally do not measure these other factors. Finally, if all other factors could be taken into account, the added-value of individual teachers could be quite small.
3. The estimation of growth has been found to vary depending on the tests used to measure it. Therefore, a teacher's value added estimate or degree of effectiveness will vary depending on how achievement is measured.

4. Value-added estimates are also affected by class size and become more unstable as class size decreases.
5. Academic achievement is not the sole goal of education. Further, the necessarily limited content of most state assessments ignores important outcomes.
6. Scores based on a single test are not adequate measures of learning.
7. Comparing a 4th grade mathematics score to a 5th grade mathematics score, for example, is problematic given that the content of the tests likely emphasizes different topics in different years. Hence, the scores are not substantively comparable.

A recent brief to policy makers jointly sponsored by the *American Education Research Association* and the *National Academy of Education*[4] highlights three major problems with the use of VAMs to yield accurate measures of teacher effectiveness:

1. *Value-added models of teacher effectiveness are highly unstable.*

Teachers' effectiveness ratings can be quite different from year to year, from class to class, and depending on the VAM used to estimate effectiveness.

Table 1: Percent of Teachers Whose Effectiveness Rankings Change

	By 1 or more Deciles	By 2 or more Deciles	By 3 or more Deciles
Across models [a]	56-80%	12-33%	0-14%
Across course [b]	85-100%	54-92%	39-54%
Across years [b]	74-93%	45-63%	19-41%

Note: [a] Depending on pair of models compared. [b] Depending on the model used.
Source: Newton, Darling-Hammond, Haertel, and Thomas (2010)[5]

2. *Teachers' value-added ratings are significantly affected by differences in the students who are assigned to them.*

The same teacher's effectiveness rating can move from the lowest category one year to the highest category the next year depending on the composition of her class, even after prior student achievement scores and class composition variables are taken into account in the model.

3. *Value-added ratings cannot disentangle the many influences on student progress.*

Student learning is influenced by many factors and many teachers and this year's achievement scores in one subject might be influenced by experiences in previous years or in other classes. The impact of these experiences can be long lasting and may manifest itself a few years later.

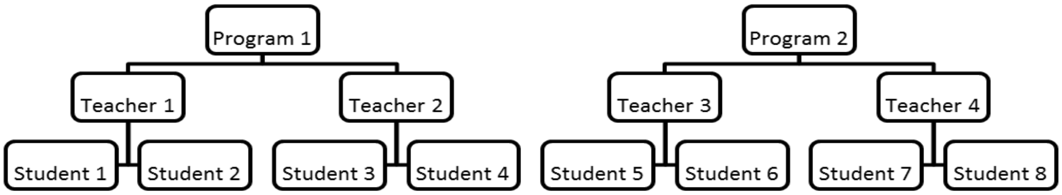
Given these problems with the use of VAMs, there is a solid consensus in the professional educational research community that these models are not appropriate as the sole basis for making important decisions about individual teachers, particularly given the unreliability and questionable validity of the ratings they generate. Thus, while in 2011 the Indiana General Assembly passed legislation requiring school corporations to develop a plan for annual performance evaluations of each certified employee that includes objectives measures of student achievement and growth, the plans must also include classroom observations and other performance indicators.

The Louisiana Value-Added Model

The Louisiana Value-Added Model is one of the models that have been developed to date. A simple presentation of this model can be understood as follows:

A model is developed that takes into account the inherent structure of test data (students and their tests scores are *nested* within their class). This sort of model, termed a multilevel model, takes into account the fact that students within a class are likely to be similar to each other since they share the same teacher, school, and possibly other demographic factors. It also allows for the inclusion of contextual factors (i.e., percentage of children in the class that are eligible for free or reduced lunch). The structure of data associated with VAMs is illustrated in Figure 1.

Figure 1. *Multilevel data structure – students nested within classes & teachers nested in teacher education programs*



This type of model is applied to student achievement test data for all students from grade 4 to 9 in a variety of content areas. Teacher value-added scores are based on the discrepancy between her students' *actual* test scores and the scores predicted by the VAM after accounting for each student's previous achievement and a number of student characteristics (i.e., disability and gifted status, free and reduced lunch, limited English proficiency, absences, prior year suspensions and expulsions) and class characteristics (i.e., percent eligible for free or reduced lunch, percent of students with disability status). The difference between each student's actual test score and their predicted score are averaged to produce a *value-added score* for each teacher. Finally, based on estimated value-added scores, new teachers are rated on the basis of whether they are performing as expected, better or worse *compared* to other teachers. New teachers are then sorted by teacher education programs from which they graduated and an average value-added/effectiveness rating is calculated for each program and content area.

Given that all students are compared to each other, regardless of the conditions in which they live or of their educational experiences, these models ensure that a certain percentage of teachers and a number of teacher education programs will always be rated as deficient –regardless of the substantive meaning of the effectiveness differences between teachers. Importantly, the substantive meanings associated with effectiveness differences are an area that has not been studied. This is quite defeating for teachers, teacher education programs, and education in general.

The Louisiana model is one among a number of models that have been developed to evaluate teacher education programs. Researchers at the University of Washington, for example, have used a slightly different model[6] that takes into account a number of variables at different levels (e.g., student background, classroom, teacher, school and district characteristics, and teacher education program indicators) to compare teacher education programs in Washington State. One difference between the Washington model and the Louisiana model is that it attempts to control for selection into teacher preparation programs by including teachers' college entrance examination scores. Other models have been developed to study programs in New York City, North Carolina, the California State University system and so on.

The issues raised earlier regarding VAMs used to evaluate individual teacher effectiveness also apply to the models used to evaluate teacher education programs and are additionally compounded by the fact that teacher education programs are one more step removed from student achievement. Additional problems are:

- Just as students are not randomly assigned to school and teachers, teacher education students are not randomly assigned to teacher education programs but rather self-select into a particular program, resulting in systematic differences between them that may not be easily controlled for by statistical modeling. Models that do not take into account this selection issue will offer results that are only descriptive and suggestive but cannot be viewed as yielding the causal impact of a teacher preparation program.
- The assumption that the teacher education program is the sole factor explaining differences between teacher effectiveness is as problematic as assuming that student learning is simply due to teaching.
- Program sizes in different institutions of higher education will have an impact on the stability and reliability of the estimates of program effectiveness. While the number of elementary teachers graduating from teacher education program may be sufficient to yield reliable estimates, secondary programs in mathematics and English Language Arts may not graduate a sufficient number of graduates every year.

Better information about differences in student performance, attitudes, interests and general well-being, and the factors that can potentially explain these differences will be helpful for educators to organize an education system that can accommodate these differences and assist students in realizing their full potential. Such information will also be useful for policy makers to articulate education policies that are not counter to understandings about how students learn and develop. Theoretically, the models hold great promise with their potential to separate non-educational factors from the effects of teachers and schools on student performance. VAMs and other models can, in theory, contribute to understanding the conditions that make learning possible assuming that these models are used in a spirit of inquiry rather than as tools used in isolation for making decisions about individuals or programs.

High-stakes decisions based solely on these models have the potential to mislead, which can harm students, teachers and education in general. If high-stakes decisions (funding, program approval, merit pay, employment, promotion) about students, teachers, schools and programs are made solely or primarily on the basis of changes in student test scores, it will further erode the meaning of education in major ways such as:

- Teaching to the test with renewed and unprecedented efforts
- Student boredom and disengagement from uninteresting teaching and from school
- Further narrowing of the curriculum to what is on the test
- De-emphasis on areas of the curriculum not tested
- Shortage of teachers for difficult teaching assignments and in tested subjects
- De-moralization and de-professionalization of teachers
- Decreased enrollment in teacher preparation programs

These consequences are, of course, not new and they are observed cyclically with each new wave of accountability mandates that put more and more pressure on schools and teachers. The new Race to the Top legislation further increases this pressure on teachers but also is targeting schools of education in an unprecedented way.

Recommendations

A value-added model considered for the evaluation of teacher education programs in Indiana should not be the sole or primary basis for making decisions about the quality of teacher education graduates or teacher education programs. At a minimum, a teacher education program evaluation system should be aligned with the teacher evaluation plans developed by the school corporations in the state. Further, consideration of the following issues will be of the highest importance to enhance the credibility of the system put in place.

1. *Gain a better understanding of teacher effects on learning.*

It is clear that teachers play an important part in student learning; however, the extent to which teachers influence this learning is unclear. Given this lack of understanding on how much a teacher can actually affect student learning, it will be important to experiment with different VAMs for a number of years to gain a perspective on the extent to which teachers can be expected to influence student learning and achievement results. Likewise, the use of VAM in teacher education program evaluation should be pilot-tested before it is fully implemented.

2. *Collaborate with stake holders and gain agreement on the best model for Indiana.*

The world of value-added modeling is complex. A number of organizations and for profit companies claim to have the most stable model. However, many of these models have not been subjected to a rigorous peer review process. In some cases where the VAM has been subjected to peer review, claims on what information the model can produce are over exaggerated. Given the complexity of VAMs it is important that policy makers, practitioners and researchers who create and study VAMs work together to create and implement a system that is disciplined and not over interpreted. Additionally, all stakeholders that will be affected by these models should be given a voice on the model's creation. It is within this type of collaborative process that the goal of creating a useful, disciplined and policy relevant educational system can be attained. Therefore, policy makers should work with all stakeholders, adherents and critics of VAM, to better understand the model's limitations and benefits before any decisions are made based on the model's findings. Specifying the models and selecting the variables that are important in the Indiana context should be a collaborative effort. The sole focus on achievement as measured by standardized test scores, for example, is a serious limitation and could have severe consequences for the life of students and teachers and the meaning of education in general. Efforts should be made to take into consideration the multiple aims of education and schooling.

3. *Understand and account for technical concerns when using VAMs and establish an audit system to monitor the quality of the data used.*

Here the old adage of "garbage in, garbage out" should be heeded. Even if the best model is chosen for Indiana that model will still be weak if the data used are of questionable reliability and validity. Technical issues such as model specification, the validity and reliability of all student growth measures, the impact of omitted variables and missing data, valid teacher and program information, the impact of small sample size have important consequences for the meaningfulness of the decisions made based on these results.

4. *Ensure transparency of the system at all times and make data and technical specifications available to researchers and stakeholders for continuous study and monitoring of the system's performance.*

Too often the exact technical specifications of the systems used and the data on which these are used are not made available for peer review, reanalysis, and further study. The credibility and meaningfulness of the systems in place can only be enhanced by ensuring transparency and collaborative responsibility for the continuous monitoring, study and improvement of these systems. This requires giving access to the data and all relevant technical information.

RAND researchers also provide a useful summary of general considerations for any system that will be using VAM for decision making purposes. The following list includes some of their recommendations and should be useful to any system that will be implementing VAMs.

- Develop databases that can support VAM estimation of teacher effects across a diverse sample of school districts or other jurisdictions
- Develop computational tools for fitting VAM that scale up to large databases and allow for extensions to the currently available models
- Link estimates of teacher effects derived from VAM with other measures of teacher effectiveness as a means of validating estimate effects
- Conduct further empirical investigation on the impact of potential sources of error in VAM estimates
- Determine the prevalence of factors that contribute to the sensitivity of estimated teacher effects
- Incorporate decision theory into VAM by working with policymakers to elicit decisions and costs associated with those decisions and by developing estimators to minimize the losses.[7]

[1] http://www.cgp.upenn.edu/ope_value.html#8

[2] Tennessee's Educational Improvement Act, 1992

[3] cite: http://www.ed.gov/news/speeches/new-approach-teacher-education-reform-and-improvement

[4] http://www.aera.net/uploadedFiles/Gov_Relations/Getting_Teacher_Evaluation_Right_summary_brief-FINAL.pdf, September 14, 2011.

[5] Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010) Value-Added Modeling of Teacher Effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18* (23). http://epaa.asu.edu/ojs/article/view/810;

[6] Goldhaber, D. & Liddle, S. (2011). The gateway to the profession: Assessing teacher preparation programs based on student achievement. Center for Education Data & Research. University of Washington, Bothell.

[7] http://www.rand.org/pubs/research_briefs/RB9050/index1.html