A Practitioner's Guide to Value Added Assessment

Edward W. Wiley
University of Colorado at Boulder

Executive Summary

Value Added Assessment (VAA) has become increasingly popular in the current context of accountability-based educational policy. Because it claims to determine how specific teachers and schools affect student learning—free of the influences of race, SES, and other contextual factors—VAA's promise as an accountability tool appears substantial. Several states and many large districts have implemented or are planning to implement accountability policies based on some form of VAA. Given this trend, it is important for practitioners and policymakers to learn more about VAA—to understand what VAA actually is and where it came from, what it can do and what it cannot do.

This guide is intended for the practitioner needing to get up to speed quickly regarding VAA. Based on a comprehensive review of current research on VAA, the guide outlines several issues that must be kept in mind when implementing a VAA-based accountability system. In addition to describing the similarities and differences among six major approaches to VAA, the guide also details several VAA-based accountability programs currently in use. Finally, the guide offers practitioners and policymakers guidance on assessing the potential of VAA for their own purposes.

A theme consistent throughout the guide is that the issues surrounding VAA-based programs are many and complex. Although different VAA approaches – each with unique strengths and weaknesses – are available for use in a given context, *all* VAA approaches share challenges that threaten the validity of teacher and school effect estimates they are designed to generate. Trade-offs and risky assumptions are required in every case, so any given model is necessarily going to be imperfect. In the context of accountability, expectations for what any VAA-based tool can reasonably accomplish should be tempered, and the use of its estimates must be judicious.

The practitioner is advised that any teacher or school effect estimated from VAA models should be taken as only that—an estimate. The weaknesses of VAA detailed in this guide render VAA inadvisable as the cornerstone of a system of teacher or school accountability. VAA-based estimates may help identify teachers who appear to be successful as well as those who appear to need assistance in improving their practice. However, until more is known about the accuracy of and tradeoffs between VAA approaches, VAA-based estimates should never serve as a single indicator of teacher effectiveness, and high stakes decisions should never be made primarily on the basis of VAA-based estimates.

A Practitioner's Guide to Value Added Assessment consists of four separate parts:

1. Background. The guide begins with a non-technical introduction to VAA that explains what VAA is and details VAA's methodological and policy roots.

2. General Issues. Once background is complete the guide surveys several general issues that the responsible user of VAA must consider. VAA systems are believed to estimate the true effect of a particular teacher or school on student learning; therefore the accuracy of assumptions underlying VAA systems is critical to the validity of VAA-based teacher and school "effects". Assumptions discussed in this section include:
    - *Attribution of teacher effects* (e.g., how can effects of two teachers be compared when their classes contain students that are incomparable demographically?)
    - *Persistence of teacher effects* (e.g., how does the influence of previous teachers carry on in subsequent years?)
    - *Rates of growth in student achievement* (e.g., given "average" teachers over a period of years, should we expect all students to show the same achievement gains every year? Or should we expect rates of achievement growth to vary across students year-to-year?)
    - *Dealing with missing data* (e.g., should teacher effects be adjusted if a student is absent on exam day? What if absent students would likely have scored lower than students who did take the test?)
    - *Use of student achievement data for teacher and school accountability* (e.g., are increases in achievement scores what we most want teachers to "produce"?)

3. Approaches to VAA. After providing this background, the guide details six major forms of VAA that attempt to measure teacher (or school) effectiveness; specific examples of VAA models currently being used by states and districts are included, with a discussion of their strengths and weaknesses. The six forms of VAA discussed include three that have been implemented as part of a formal accountability system and three that have been proposed but have yet to be implemented for accountability purposes:
    - Simple *gain score* models (e.g., Texas Growth Index)
    - *Covariate adjustment* models (e.g., Dallas Value Added Assessment System or "DVAAS")
    - William Sanders's *layered* models (e.g., Educational Value Added Assessment System or "EVAAS"; Tennessee Value Added Assessment System or "TVAAS")
    - *Cross-classified* models
    - RAND's *persistence* model
    - Todd and Wolpin's *cumulative within-child mixed-effects* model.

4. Selecting a VAA-based approach. Finally, the guide offers practitioners and policymakers guidance on assessing the potential of VAA for their own purposes.
    - Among *multiple*-wave approaches (that is, those that can model more than two years of data), RAND's *persistence* model is preferred to other alternatives due

to its flexibility. Approaches that assume undiminished teacher effects (which research has shown to be questionable), are not recommended.

- Among *single-wave* approaches, the covariate adjustment model is preferred to the gain score model as the covariate adjustment model also does not assume undiminished persistence of teacher effects.
- Regardless of which model seems best suited to a particular context, general issues relating to the validity of VAA estimates must also be taken into consideration. No set of adjustments can fully compensate for the lack of randomization that would support causal claims about the effects of teachers.
- In addition to the problem of lack of randomization, important questions remain about whether student achievement scores are appropriate measures of teacher effects; whether achievement tests are appropriately designed; whether and how assessment errors may affect estimates; and when assessments are best administered. Although these are crucial and complex questions, no VAA approach yet takes them into account.
- Expectations for what any VAA-based tool can reasonably accomplish should be tempered, and the use of its estimates should be judicious. VAA-based estimates may help identify teachers who appear to be successful as well as those who appear to need assistance in improving their practice. However, until more is known about how to improve the accuracy of VAA approaches, VAA-based estimates should never serve as a single indicator of teacher effectiveness, and high stakes decisions should never be made primarily on the basis of VAA-based estimates.

A Practitioner's Guide to Value Added Assessment

Edward W. Wiley
University of Colorado at Boulder

Introduction

Incredibly, you can walk into almost any school in America, go down the hall to the first couple of classrooms you find, look at the teachers inside, and realize this: nobody, not the principal, not the parents, not the students, not even the teachers themselves, actually knows how effective those teachers are in helping their students learn. They probably have an opinion, maybe even some anecdotal evidence. But in terms of accurate, verifiable information about how effective individual teachers are at helping each of their students learn and make progress from the beginning of the school year to the end? In the vast majority of schools, nobody knows.

-- Kevin Carey, EdTrust[1]

[Through value-added assessment] educational influences on the rate of student progress can be partitioned from exogenous factors (if not completely, then nearly so) allowing an objective measure of the influence of the district, school and teacher on the rate of academic progress.

-- Bill Sanders[2]

All models are wrong but some are useful.

-- G. E. P. Box[3]

The search for "accurate, verifiable information" about the effectiveness of teachers and schools has long frustrated educational researchers and practitioners. However, many now believe that Value Added Assessment (VAA) holds the long-awaited solution to this problem. Because it claims to determine how specific teachers and schools affect student learning—free of the influences of race, socio-economic status (SES), and other contextual factors—VAA has become very popular in the current context of accountability-based educational policy. Its promise as an accountability tool

appears to be substantial. In fact, several states and many large districts already have implemented or are planning to implement policies based on some form of VAA. Given this trend, it is important to learn more about VAA—to understand what VAA actually is, where it came from, and what it can and cannot do.

This discussion is intended to serve as a user's guide for practitioners and policymakers considering implementing a VAA-based accountability system. Informed by a growing body of research on VAA[4] and by an Educational Testing Service primer,[5] this guide reviews critical validity issues and provides a detailed look into the particular variants of VAA currently in use. The final section offers some advice for those interested in exploring VAA feasibility for their own purposes.

Value Added Assessment: Background

The objective of VAA is straightforward: to attribute changes in student achievement to sources responsible for those changes—most commonly teachers and schools. The output of VAA is an estimated teacher (or school) "effect" —a numerical measure of the *effectiveness* of a particular teacher or school. In its most basic form, VAA is based simply on the calculation of year-to-year changes in students' test scores; more complicated forms of VAA incorporate statistical techniques to account for such factors as differences in student characteristics and persisting effects of previous teachers and schools.

Although VAA is a form of "growth" or "longitudinal" model (that is, a model based on changes in assessment scores over time), the terms "value added assessment" and "growth model" are not interchangeable. VAA is best considered a type of growth model that tracks scores over time for individual students in order to estimate how much

change can be attributed to teachers or schools. Other types of growth models exhibit significant differences from VAA. These would include, for example, the current model for "Safe Harbor" under No Child Left Behind",[6] the growth objectives under California's Academic Performance Index,[7] and the "Required Improvement" metric of the Texas Accountability System.[8] Each of these programs measures year-to-year change by comparing two successive cohorts; however, a significant difference is that cohorts in these programs do not necessarily include the same students. In contrast, VAA tracks individual students. Other programs, such as North Carolina's "ABC's of Public Education",[9] and Arizona's "Measures of Academic Progress",[10] do track year-to-year change for individual students; however, these programs base their estimates of student performance on targets for average growth (in the case of North Carolina) or on the percent of students achieving targeted growth (in the case of Arizona). Because such programs focus on growth itself rather than on factors that contribute to it, these too do not strictly qualify as examples of VAA. In essence, the distinguishing characteristics of any VAA model are that: 1) it studies change in the performance of individual students, and 2) it seeks to determine to what extent changes in student performance may be attributed to particular schools and teachers.

VAA's methodological roots lie both in econometrics and educational statistics. Economists have long employed "production function" models to describe mathematically how a firm creates output from its inputs—how it uses resources and procedures to produce a product. The production function measures productivity (that is, the value created by) a specific collection of inputs. Valuable inputs are those that are more productive—they provide greater output per unit, or greater quality per unit, or

more of some other valued characteristic. For example, productivity measures provide a method of sorting and discriminating among workers (input) based on the quality and/or quantity of their work (output).

Economists interested in education have applied the input/output model to estimate how various factors affect the outcomes of schooling.[11] The question central to analyses via "Education Production Functions" (EPF) is to what extent changes in student performance, or output (e.g., mathematics achievement scores), can be attributed to particular inputs (especially teachers and schools, or perhaps educational reforms) "received" by the student over a specified period of time. EPF estimates of the effects of a particular teacher on student learning are analogous to the estimated effects of a particular worker's efforts on a firm's output.

Economists have not been alone in exploring measures of teacher and school effectiveness; educational statisticians have also developed models to address the same cause and effect questions. Methods for longitudinal analysis of student assessment data share a history similar to that of education production functions. Early models were based on either simple year-to-year changes in scores or predictions of current-year scores using previous year scores. As shortcomings of these approaches became apparent,[12] they gave way to more complex approaches. Today educational statisticians employ complex statistical models (known variously as "hierarchical linear models",[13] "multilevel models",[14] or "random effects models",[15]) to account for the "nested" structure of educational settings—groups of students belong to a particular teacher, groups of teachers teach in a particular school, and groups of schools make up a particular district.

VAA's methodological predecessors may have never emerged as prominent policy tools had it not been for the rise of standards-based accountability. The 2002 reauthorization of the Elementary and Secondary Education Act, *No Child Left Behind,*[16] gave a push to an assessment-based accountability movement that had already gained momentum in many states. VAA's benefits appear perfectly matched to the requirements of such accountability systems. In the past year, the VAA movement has picked up even greater momentum, as states have requested increased flexibility to allow Adequate Yearly Progress (AYP) plans based on growth models. U.S. Secretary of Education Margaret Spellings has responded to the increasing enthusiasm for growth models by announcing support for ten statewide pilot programs that incorporate growth models into AYP.[17] If shown to provide valid inferences, AYP programs based on growth models (and possibly VAA) will likely see even more widespread adoption in the coming years.

VAA is increasingly a darling of policymakers and is gaining support among some scholars. Accountability advocates attracted to its promises of facilitating apples-to-apples comparisons of the effectiveness of specific schools or teachers—free of context effects that have traditionally made such comparisons problematic—continue to push for its adoption at district and state levels. VAA appears to be here to stay—at least for the foreseeable future.

Value Added Assessment: General Issues

The term "Value Added Assessment" actually encompasses several different statistical models. Varying substantially in complexity and underlying assumptions, these approaches all satisfy the same necessary condition: they link changes in individual student achievement to particular teachers or schools.

Much recent research has focused on methodological and other substantive issues surrounding VAA as an accountability tool. The most comprehensive discussion of such research to date is provided by McCaffrey et al.;[18] also, a non-technical review of this discussion was written by Braun.[19] The next section draws from each of these works to provide a brief overview of issues concerning the validity of VAA-based accountability models.

Attribution of Teacher Effects

Braun[20] describes what he terms the "fundamental concern" about VAA: whether VAA systems can in fact reliably estimate the effect of a particular teacher or group of teachers (or school/s) on student learning: "…the reader [of typical VAA studies] is invited to treat [teacher effect estimates] as if, in fact, they unambiguously capture the relative contributions of different teachers to student learning." A major issue here is that cause and effect research generally requires that subjects (students) be randomly placed into treatment groups (teacher classrooms). However, randomly grouping students into classes and randomly assigning teachers to those classes are unlikely in school systems; teachers are often assigned classes based on seniority, and various forms of ability grouping and tracking are common. Moreover, schools within a given district, and districts within a given state, enroll students varying greatly in prior achievement, skills, and knowledge. In the absence of being able to randomly group students into classes, researchers typically make statistical adjustments for identifiable and quantifiable "irregular" circumstances (such as an imbalance in student ability across different classes) that would cause teacher effect estimates to be inaccurate. However, even when statistical adjustment is made, it is unlikely to make enough difference to support

cause/effect claims about teacher performance. As Braun notes: "It is impossible…to document and model all such irregular circumstances; yet they may well influence, directly or indirectly, the answers we seek nearly as much as what the teacher actually does in the classroom."[21]

The heart of this problem is that the inability to randomly group students and teachers can result in factors other than teacher effectiveness influencing VAA estimates. Because students are not randomly assigned to teachers, and teachers are not randomly assigned to classes, students assigned to one teacher may differ significantly in capability and family context[22] from those assigned to another teacher. Class characteristics may differ as well; for example, as Todd and Wolpin point out, "if a popular teacher has higher enrollments and class-size is omitted from the specification, then the estimated teacher effect will include the impact of class size on performance."[23] Or, as Braun notes:

> [the] implicit assumption that [teachers being compared] all have been assigned similar academic goals for their classes and have equivalent resources…flies in the face of the reality that tracking is endemic in schools, particularly in middle schools and above. Students in different classes may be exposed to different material and assigned different end-of-year targets. These differences will influence the estimates of teacher effects.[24]

Similarly, schools may differ in policies, resources, and contexts: "…different schools in the same district may be employing different curricula or following different reform strategies."[25] And parents too may influence student performance by promoting different activities (such as more reading at home or after-school tutoring).

Even if the ideal situation—randomly grouping teachers and students—were possible, the small number of students a teacher works with each year can influence a teacher's estimated effects. Annual changes in class composition can produce year-to-year volatility that makes the estimated teacher effect in any given year atypical.[26] Again, an example provided by Braun:[27]

> With a relatively small number of students contributing to the estimated effect for a particular teacher, the averaging power of randomization can't work for all teachers in a given year. Suppose, for example, that there are a small number of truly disruptive students in a cohort. While all teachers may have an equal chance of finding one (or more) of those students in their class each year, only a few actually will — with potentially deleterious impact on the academic growth of the class in that year. The bottom line is that even if teachers and students come together in more or less random ways, estimated teacher effects can be quite variable from year to year.

As this section demonstrates, many factors other than teacher performance may be reflected in estimates of teacher effectiveness; they are difficult to disentangle from true differences across teachers. Striving to make estimates more reliable, researchers have devised various statistical methods to accommodate other factors; differences in these accommodations are one of the main ways various VAA approaches differ from each other.

Persistence of Teacher Effects

VAA approaches also differ in the assumptions they make regarding the persistence of teacher effects in the years after students have moved to other classes. Some models assume that a teacher's effect extends undiminished into the future. Other models make no such assumption; rather, the persistence of teacher effects (be it constant or diminishing over time) is estimated by the data. This assumption affects the degree to which changes in student scores are attributable alternatively to current and previous teachers; therefore it is likely to have significant ramifications on teacher effect estimates. For example, if prior teachers continue to be credited for a student's current performance, the effect estimated for a current teacher whose students previously had teachers with large positive effects will tend to be artificially lowered.

For example, let's say that teacher effects truly do diminish over time, at a rate of 20% of the initial effect per year. Consider a class of students that started with a 2nd-grade average score of 100, then had a 3rd grade teacher with an effect of 5 points, followed by a 4th grade teacher who had an effect of 3 points. That class would, on average, score 107 on the 4th grade test:

100	(Average at end of 2nd grade)
+ 4	(Contribution of 3rd-grade teacher to 4th grade test; the original effect of 5 has diminished by 20% over the single year since having that teacher)
+ 3	(Contribution of 4th-grade teacher)
=107	(Average score on 4th grade test)

In this case, if teacher effects are assumed (erroneously) to persist undiminished (rather than diminish 20% each year), the effect for the 4th grade teacher would be underestimated:

107	(Average score on 4th grade test)
-100	(Average at end of 2nd grade)
-5	(Estimated effect of 3rd grade teacher to 4th grade test; original effect of 5 is assumed to have persist undiminished over the single year since having that teacher)
= 2	(Estimated effect of 4th-grade teacher; actual effect is +3)

In this example the 4th grade teacher's effect was artificially lower than it should have been, because the 3rd grade teacher's positive effect was assumed to persist undiminished. In this case, because of the assumption that teacher effects persist undiminished, a previous teacher with a positive effect was given credit for achievement effects that should have been attributed to a future teacher.

Nature of growth in student achievement

Another way in which VAA approaches differ is in their assumptions about growth in student achievement. Given only "average" teachers over a period of years, should we expect student achievement to grow at a constant (linear) annual rate? Or should we expect growth to vary among individual students and over different years? Different assumptions about student growth lead to different VAA models, which in turn are likely to yield significantly different estimates of teacher effectiveness.

Missing Data

Missing data is a major challenge to any statistical estimation. VAA requires not only consistent records of student performance but also reliable class rosters, so that individual students can be matched to individual teachers. Data may be missing because of unreliable records, but may also be missing because of other reasons. For example, absenteeism certainly causes missing data. Exemption of certain students from testing, due to parental requests of waivers or identification of students for whom the test is believed to be inappropriate (e.g., due to limited English proficiency), also causes test data to be missing for certain students. Another, more troubling cause of missing data may be related to "gaming the system" due to the high-stakes nature of assessment scores – for example, the threat of sanctions may give a teacher or principal an incentive to push a kid out of the school or at least encourage the kid to skip out on the test.

McCaffrey et al.[28] describe in detail several potential impacts of missing data. Unfortunately there is no simple, straightforward way to deal with the challenge. Some variants of VAA simply exclude students for whom complete data are not available. However, if data are missing not randomly but for some systematic reason, teacher effects could be skewed. For example, students who fail to take achievement tests—and whose data would therefore be missing —tend to perform less well than students who do take the tests. If a number of low-performing students were simply excluded from the study, the teacher's effectiveness with low performers would not be fairly reflected.

Rather than exclude subjects, other VAA approaches make assumptions about the patterns of missing data. Based on these assumptions, researchers generate missing data so that the performance of all students can be included in analyses. Rather obviously,

however, if the assumptions are not correct and the data generated are not accurate, the teacher effects yielded by these approaches may also be skewed.

Issues in Using Student Achievement Data

Significant research has focused on the promise and perils of using student achievement data as an outcome—as a good indicator of teacher effectiveness—in accountability models. Several questions about VAA have been raised in this area. Are increases in achievement scores what we most want teachers to "produce"? Are tests at each grade level reliably appropriate for the grade level? Do they assume a reasonable and feasible amount of growth from year to year? Are scores across different grades comparable? When different versions of the same test are used, can we be sure they measure exactly the same thing in exactly the same way? When are the appropriate times to measure growth—fall to spring, or spring to spring? Such questions are crucial to interpreting VAA teacher effect estimates;[29] and, while all of them have significant implications for the accuracy of these estimates, no current VAA approaches explicitly take them into account.[30]

Approaches to Value Added Assessment

Although major issues remain unresolved, several VAA models have nevertheless been developed and implemented. Following is an overview of six alternative models, including each model's unique characteristics[31] as well as its strengths and weaknesses. Specific examples of models already in use are also discussed. In total, this overview provides a practitioner's guide to the current status of VAA as a tool for accountability systems. Although models are described as they apply to assessing the effects of

teachers, they might just as easily be used to estimate the effects of schools or other units of analysis.

To ensure that this guide is accessible to a broad audience, technical exposition has been kept to a minimum. Therefore, the complex statistical underpinnings of each model do not appear here; readers interested in such detail are directed to McCaffrey et al.[32] and Todd and Wolpin.[33] The six VAA forms presented in the next section, and the primary dimensions on which they differ, are summarized in Table 1. As the table indicates, a primary characteristic of all models is whether they are "single wave," based on only two measurements (perhaps a student's fall and spring reading achievement scores), or "multiple wave," based on more than two measurements (perhaps a student's reading achievement scores for grades 2, 4 and 6).

Table 1. Types of VAA Models

		VAA Model						
		Single-wave models (two sets of measurements)			**Multiple-wave models** (multiple sets of measurements)			
		Gain Score	**Covariate Adjustment**		**Layered**	**Cross-Classified**	**Persistence**	**Cumulative Within-Child**
Issue								
Years of Data		Maximum of 2	Maximum of 2		Multiple	Multiple	Multiple	Multiple
Content Areas		Maximum of 1	Maximum of 1		Multiple	Multiple	Multiple	Multiple
Student Cohorts		Maximum of 1	Maximum of 1		Multiple	Multiple	Multiple	Multiple
Persistence of Teacher Effects (How does the influence of previous teachers carry on in subsequent years?)		Teacher effects persist undiminished	Teacher effects may diminish over time		Teacher effects persist undiminished	Teacher effects persist undiminished	Teacher effects may diminish over time	Teacher effects may diminish over time
Rates of Growth in Student Achievement (Given "average" teachers over a period of years, does student achievement grow at the same average rate every year? Or are rates of achievement growth unique year-to-year?)		No assumption made	No assumption made		No assumption made	Annual growth in student achievement assumed to be constant (linear)	No assumption made	No assumption made (though possible)
Treatment of missing data (Should teacher effects be adjusted if a student is absent on exam day?)		Students with missing data excluded from analysis	Students with missing data excluded from analysis		Students with missing data included using projected ("imputed") scores	Students with missing data included using projected ("imputed") scores	Students with missing data included using projected ("imputed") scores	Students with missing data included using projected ("imputed") scores
Inclusion of student demographic characteristics (e.g., ethnicity, poverty, ELL status)		Possible	Possible		No	Possible	Possible	Yes
Consideration of student's previous educational experiences (e.g., participation in preschool programs or reduced class sizes in primary grades)		Possible	Possible		No	Possible	Possible	Yes
Consideration of family inputs (e.g., amount of reading done with parents at home; participation in after-school tutoring programs)		No	No		No	Possible	Possible	Yes
Consideration of child inherited attributes (a child's inherited intellectual ability)		No	No		No	Possible	Possible	Yes

The simple "gain score model"

A student's year-to-year change in achievement scores provides the basis for VAA in its most simple form. The "gain score model" links the one-year gain in a student's score to that student's teacher. For each student, last year's score is subtracted from this year's score to measure the student's change in achievement while in a particular teacher's classroom. Gains for all of one teacher's students are averaged, and the average is then compared to that of other teachers' students. For example, one teacher's average gain score might be compared to the average gain score of all other teachers in the same building, or in the same district. The "teacher effect" is the difference between an individual teacher's gain score and that of the comparison group. This model can include statistical adjustments for student characteristics; however, the basic idea—that the best teachers are those whose students show the biggest year-to-year gains in achievement tests—remains the same.

The following simple example describes how the gain score model works; in most practical cases the statistical mechanics are more complex (and comparisons are made at the district rather than school level), but the general idea is consistent. Consider a (very small) school with three 4th grade teachers, each of whom has only three students. Scores for these students on the state math assessment are as follows:

Teacher	Student	4th grade Math Score	3rd grade Math Score	Gain, 3rd to 4th	Average Gain, Teacher	School Average Gain	Teacher Effect
1	1	35	25	10	8.67	4.44	4.22
	2	33	27	6			
	3	39	29	10			
2	4	33	31	2	2.67	4.44	-1.78
	5	35	33	2			
	6	39	35	4			
3	7	43	41	2	2.00	4.44	-2.44
	8	47	43	4			
	9	45	45	0			

In this example all three teachers have positive gains; each one's students have (on average) gained in math knowledge from the grade 3 assessment to the grade 4 assessment. Teacher 1 has the greatest "effect," however, with a classroom gain more than 4 points larger than the school average. Even though their students demonstrated positive gains, teachers 2 and 3 had negative teacher effects; their gains were about 2 points below the school average.

A primary strength of the gain score model is that it is intuitively easy to understand; better teachers are those whose students show greater gains in achievement scores. Additional advantages to the gain score model are that it requires only two years of data in a single subject, and that implementation is straightforward with commonly available software (such as Microsoft Excel or SPSS).

The gain score model is not without its shortcomings, however. First, it considers only two years of data; therefore, estimating teacher effects over five years would require four separate analyses. Second, it does not consider student achievement in previous

years, which more complex models take into account. In addition, gain score models assume that the effects of a student's previous educational experiences (including previous teachers) persist into the future undiminished. To accept a gain score estimate of teacher effects as valid, then, means accepting that the effects of a student's prior teachers never fade, and that a teacher's effectiveness is best measured by the achievement scores of her most recent students.

Another shortcoming is that the gain score model does not consider where students started. In other words, the gain score model treats gains as similarly meaningful no matter where they appear on a scale. For example, on a 100-point test a student's gain from a score of 15 to 20 appears the same as a student's gain from 85 to 90; both students show a gain of 5. These two gains may not be comparable instructionally, but the gain score model treats them as the same.

Finally, the typical gain score model does not use information from students for whom data is missing. Students without complete information are excluded from analysis. If the nature of missing data is not random—if, for example, scores are missing primarily for high-mobility students who tend to have lower achievement scores—the teacher effects estimate could be skewed, as noted earlier.

Characteristics of the gain score model are summarized in Figure 1.

Figure 1.

Gain Score Model

Outcomes modeled:

- Individual assessment scores across a single period (two score comparison)
- Single Subject
- Single Cohort

Adjustments: Student and School characteristics

What differentiates it from covariate adjustment models

- Assumption that teacher effects persist undiminished

Key Strengths

- Simplicity; easy to understand and explain
- Straightforward implementation

Key Shortcomings

- No consideration of information from previous years
- Exclusion of students with missing data (may skew teacher effects)
- No consideration of point at which students start along the developmental scale
- Assumption of undiminished teacher effects (may not be reliable)
- Omission of statistical adjustments for student ability (may skew teacher effects)
- No accommodation for problems associated with using student assessment scores

Accountability programs based on the gain score model:[34] Texas Growth Index

The Gain Score Model in Practice: The Texas Growth Index[35]

The Texas Growth Index (TGI) provides an estimate of a student's academic growth based on scores from the Texas Assessment of Knowledge and Skills (TAKS) in two consecutive grades. Texas uses the TGI primarily to identify schools eligible for state performance awards (the Gold Performance Acknowledgment for Comparable Improvement in Reading/ELA and Mathematics); TGI also plays a role in some alternative accountability systems.

For each student with two years of TAKS scores, an "expected" current-year score is calculated by multiplying the previous year's score by some amount of growth that the state has established as a target. The expected score is subtracted from the student's actual score and adjusted to account for differences across grades. The result is a student's TGI score. If TGI is zero (as would be the case if the expected grade and the actual grade were the same), the inference is that one year's growth has occurred. Higher scores indicate more rapid learning. Student TGI scores in reading and mathematics are averaged for the school.

School TGI scores are ranked relative to average TGIs of a comparison group—created uniquely for each school—of 40 other Texas schools most similar demographically (based on ethnicity, economic disadvantage, limited English proficiency, and student mobility). A school's Comparable Improvement score comes from a comparison of its students' average growth with the average student growth in comparison schools. It is measured in quartiles; a "quartile 1" school would rank in the top 25% of schools in its comparison group. Schools receive separate quartile rankings for reading and for mathematics.

TGI scores are assumed to reflect school-level contributions to student growth. These scores should be interpreted with caution, however, as the TGI program exhibits many of the shortcomings common to gain score models (described above and listed in Figure 1). Like other VAA-based approaches, TGI relies on student achievement data, which has been questioned as an appropriate indicator, and it does not provide for random selection of students. TGI also considers only two years of data, disregarding potentially useful information from previous years. The system excludes from

consideration those students with missing data. Finally, student growth targets are constant, no matter where students begin on a developmental scale; the TGI program expects that student growth should be the same whether a student started out very high or very low.

As Texas state policy-makers appear to realize, these shortcomings create uncertainty that makes TGI scores unsuitable for high stakes decisions. At this point, Texas uses TGI scores primarily for recognizing schools for commendable performance, as noted above. The percentage of students with TGI scores greater than zero (denoting positive growth) is also used as one of many indicators in alternative education accountability (AEA) systems designed specifically for charter schools and for schools with nontraditional programs.

The "covariate adjustment model"

A second basic type of VAA based on change over a single period is the covariate adjustment model, which has been most prominent in the EPF literature of educational economists.[36] This model is similar to the gain score model in linking achievement changes to current teachers only and in allowing for adjustments based on student characteristics.

The primary difference between the covariate adjustment model and the gain score model is in how the two models treat the effects of previous teachers. As noted above, the gain score model assumes these effects are permanent and unchanging. In contrast, the covariate adjustment model makes no such assumption; rather, persistence of teacher effects is estimated. This allows for effects of previous teachers to change (generally, to diminish) as time passes.

Outside of this difference, the covariate adjustment model shares the strengths and weaknesses of the gain score model. It is easy to understand and implement, requiring only two years of data in a single subject (a "single wave"); it excludes students with missing data; and it treats gains similarly no matter where they occur along a developmental scale. Because it is limited to a single wave of two assessment scores, separate analyses are necessary for multiple cohorts and subjects.

Characteristics of the covariate adjustment model are summarized in Figure 2.

Figure 2.

Covariate Adjustment Models

Outcomes modeled:

- Individual assessment scores across a single period (two score comparison)
- Single Subject
- Single Cohort

Adjustments: Student and School characteristics

What differentiates it from gain score models

- Previous teacher effects modeled by data and allowed to diminish

Key Strengths

- Simplicity; easy to understand
- Straightforward implementation
- Allowance for effects of prior educational experiences, including teachers

Key Shortcomings

- No consideration of information from previous years
- Exclusion of students with missing data (may skew teacher effects)
- No consideration of point at which students start along the developmental scale
- Omission of statistical adjustments for student ability (may skew teacher effects)
- No accommodation for problems associated with using student assessment scores

Accountability programs based on the covariate adjustment model: Dallas Value Added Assessment System (DVAAS)

The Covariate Adjustment Model in Practice: Dallas Value Added Assessment System (DVAAS)[37]

The Dallas Value Added Assessment System (DVAAS) has been in place since the early 1990's, ranking it among the oldest of VAA-based accountability programs. Its name is similar to the most famous VAA-based approach—the Tennessee Value Added Assessment System, or TVAAS; however, the two should not be confused. The DVAAS

differs significantly from its more famous counterpart, as will become clear in the later discussion of TVAAS.

The DVAAS model allows for a variety of student- and school-level factors that affect student performance. The model begins with scores for student growth from a prior period, and then adjusts them to allow for student characteristics such as ethnicity, gender, language proficiency, and socio-economic status. Scores are adjusted a second time for school-level factors such as student mobility, crowding, overall socio-economic status, and percentage minority. Teacher and school effects are estimated based on averages of student scores once all adjustments have been made.

DVAAS uses different calculations each year, with only two time points included in each set of calculations. Separate estimates for reading and math are made across all grades, as well as for writing, science and social studies in selected grades. Students missing achievement data are eliminated from some analyses, as are students with high absenteeism, the apparent assumption being that students' absences should not be attributed to their teachers.

DVAAS estimates of teacher and school effects should be used with caution for several reasons. Like other VAA-based approaches, DVAAS is faced with the challenges of using student achievement data as its outcome measure. Missing data is also a problem for DVAAS; the exclusion of high-mobility and frequently absent students could easily skew results, as these students tend to perform less well on achievement measures than other students.

However, DVAAS does include student- and school-level characteristics, which may mitigate some of the issues caused by non-random grouping of teachers and

students. How well the model actually accommodates significant differences among students and schools depends on the specific calculations used—although no statistical representation can perfectly capture teacher and student characteristics under non-random assignment, and so the validity of estimated effects may remain questionable.

A final weakness of DVAAS is the limited capacity of covariate adjustment models. Students typically take assessments in several subjects across multiple years, and information provided by the combination of these many assessments is potentially useful in estimating teacher and school effects. However, DVAAS must ignore this wealth of information because covariate adjustment models are single wave—limited to only two student achievement scores.

Such concerns argue for caution in interpreting DVAAS-generated estimates of teacher and school effects, and Dallas wisely does not base high stakes decisions solely on these estimates. Instead, they are used primarily to help design teacher and school improvement plans. Although they may be considered in personnel decisions, estimates ultimately play only a small part in such decisions.

The "layered" model

The gain score model and covariate adjustment model are similar in that they are "single wave"—that is, they consider changes across a single period characterized by only two points of measurement (fall and spring reading scores, for example). All other models (including the "layered model," which will be discussed momentarily) are "multiple wave"; that is, they consider more than two sets of measurements. Also referred to as "multivariate"[38] or "cumulative" models,"[39] such approaches are more complex in that they use multiple assessments to estimate how multiple teachers across

multiple years affect student achievement. Multiple wave models can, for example, provide teacher effects estimates based on both reading and math assessments for several groups of students over several years.

The most well-known multiple-wave model is a "layered" model, the Tennessee Value-Added Assessment System (TVAAS). William Sanders, the father of the TVAAS, has been so prominent a champion of VAA that many refer to the TVAAS—and even VAA in general—as the "Sanders Model." Designed specifically to measure educational outcomes in Tennessee, the layered approach now underlies statistical models used by the many clients of EVAAS (Educational Value-Added Assessment System), the commercial version of TVAAS that Dr. Sanders manages for the SAS Institute, Inc.[40]

As a multiple wave model, the layered model uses several years of data to estimate teacher effects. Like the gain score model, however, it assumes that teacher effects persist undiminished in subsequent years: a particular 3rd grade teacher has the same impact on a student's 3rd grade assessment as she does on the student's 6th grade assessment. Together, these characteristics mean that a layered model not only assumes that teacher effects persist undiminished, but that they can be estimated over multiple years. While these may be problematic assumptions, the layered model does move beyond the disconnected year-to-year estimates of gain score models.

Nothing would prohibit the layered model from including student background factors—but TVAAS does not include them. In the words of its developers, "TVAAS uses a highly parsimonious model that omits controls for SES, demographic, or other factors that influence achievement."[41] Student characteristics are not considered in any

way because the model assumes student scores in prior years adequately reflect student characteristics. According to SAS:

> Value-added assessment eliminates the possibility of a distorted view of effective schooling by following the progress of individual students. With SAS EVAAS methodology, each student serves as his or her own control, creating a level playing field and eliminating the need to adjust for race, poverty, or other socioeconomic factors. This innovative approach ensures that the results are fair to both students and educators."[42]

However, not everyone agrees that this is the best way to deal with the complexity of educational contexts; the issue of whether statistical adjustments for student characteristics need to be included in VAA models is far from being resolved.[43]

The layered model teacher effect is the difference between a specific teacher's average gain and the average gain of all district teachers. Because the effect is calculated as an average across all of a teacher's students, factors idiosyncratic to any given student do not enter into the teacher effect. No adjustments are made on the basis of student characteristics or previous educational experiences (though nothing prohibits such adjustments), and effects of former teachers persist undiminished over time. Characteristics of the layered model are summarized in Figure 3.

Figure 3.

Layered Model

Outcomes modeled:

- Individual assessment scores across multiple time periods (typically five years of assessment scores)
- Multiple subjects possible
- Multiple cohorts possible

Adjustments: None

What differentiates it from other multiple-wave models

- Exclusion of adjustments for student or school characteristics
- Assumption that teacher effects persist undiminished
- No assumption about student achievement growth

Key Strengths

- Inclusion of information from previous years
- Capability to include students with missing data (though missing scores must be estimated)
- No requirement for information on student or school characteristics
- Uses multiple years/subjects rather than student controls

Key Shortcomings

- Complexity; difficult to understand and explain
- Required specialized (commercial) software
- Projected scores for students with missing data (may skew teacher effects)
- No consideration of point at which students start along the developmental scale
- Assumption of undiminished teacher effects (may not be reliable)
- Omission of statistical adjustments for student ability (may skew teacher effects)
- No accommodation for problems associated with using student assessment scores

Accountability programs based on the layered model: Educational Value Added Assessment System (EVAAS)

The Layered Model in Practice: Educational Value Added Assessment System (EVAAS)[44]

With the adoption of TVAAS, Tennessee became the first state to adopt a statewide accountability system based on VAA. The TVASS was the first generation of the now widely available layered model known as EVAAS (the Educational Value Added Assessment System).

EVAAS estimates teacher effects using student assessment scores across multiple years and subjects (Tennessee, for example, uses five years of assessment data across five different subjects). To account for the relationships between these multiple assessments, the EVAAS model uses complex statistical processes which embed two key assumptions. The first is that a student's test score reflects the effects of both the current teacher and all previous teachers; the second is that a student's test score reflects (or captures, for purposes of statistical modeling) the student's personal characteristics. To estimate the effect of a specific teacher, then, the model makes statistical adjustments to student scores in order to allow for the effects of previous teachers. (As noted above, no adjustments are made for student characteristics because proponents believe such adjustments are unnecessary.) The average of the adjusted student scores is compared to the district's average; the difference between the teacher's average and district's average indicates the teacher's estimated effectiveness. Estimates are typically averaged over three years for each individual teacher.

EVAAS is the most widely used approach to VAA-based accountability. As is true for other models, however, its results should be used with caution. Several issues—some common among VAA approaches and some specific to EVAAS—may undermine

the validity of estimated teacher effects. First and most obviously, EVAAS shares with other VAA-based approaches the multiple issues surrounding the use of student achievement data as an outcome measure. As is also true for other approaches, EVAAS must deal with the challenge of missing data; however, the fact that EVAAS uses test scores across multiple years and subjects may mean that its estimates are less affected by missing data than estimates from models using fewer measurements. In addition, EVAAS reduces the year-to-year volatility of single wave teacher effect estimates by averaging estimates from three years to arrive at a single teacher-effect score.

Unique to EVAAS is the exclusion of student and school characteristics, which may well skew its estimates. EVAAS developers claim that the exclusion of these characteristics makes little difference,[45] although they note that differences may be bigger in districts characterized by greater stratification among teachers and students. This issue is far from being resolved, and many studies have pointed to the potential impact of omitting important student- and school-level characteristics.[46]

Another issue which may affect the validity of EVAAS estimates is the assumption that previous teachers' effects persist undiminished throughout a student's education. This assumption has been questioned, and empirical analyses have suggested that teacher effects do actually diminish as students progress through school.[47] If this is indeed the case, contrary to the EVAAS assumption, teacher effects will be predictably skewed. The estimated effects of teachers whose students were previously taught by teachers with large positive effects will tend to be artificially lowered (because prior teachers continue to be credited for a student's current performance); the estimated effects of teachers whose students were previously taught by teachers with negative

effects will tend to be artificially raised (because prior teachers continue to be blamed for a student's current performance).

It is important that any user of EVAAS understand these concerns about the validity of EVAAS scores. At the very least, due to the uncertainty around these issues, EVAAS scores should not be used as the sole basis for high-stakes decisions.

Additional Layered Models in Practice: Programs Based on EVAAS

EVAAS-based programs have been implemented or are being considered in over 300 school districts in 21 states.[48] These include: 65 districts in Colorado;[49] districts in Seattle[50] and in Minneapolis;[51] and districts partnering with the Iowa Association of School Boards.[52] The New York State School Boards Association (NYSSBA) is also planning an EVASS pilot partnership.[53]

In addition, Ohio plans to include a layered VAA system as part of its school performance index by 2007. A pilot program, known as "Project SOAR" (Schools' On-line Achievement Reports), "is assisting districts in their efforts to focus instruction to improve performance, raise achievement levels and help students meet Ohio's academic content standards".[54] The Project SOAR model will track districts, schools, grades, classrooms, and individual students primarily to provide performance information that will help planning for improvement.[55] Not surprisingly, since it is based primarily on EVAAS, the Ohio model is very similar to Tennessee's. The Ohio Partnership for Accountability also uses value-added assessment to estimate the effects of new teachers, an effort that provides information about teacher preparation programs and practices.[56]

These examples are each similar to the EVAAS, and therefore the same caution about using results as a sole base for high-stakes decisions applies. Some implementers

already make allowances for EVAAS' possible weaknesses. The Pennsylvania Value Added Assessment System (PVAAS)[57] is a case in point. The state has taken two precautions in using effectiveness estimates from the system. First, because questions remain concerning the validity of VAA estimates, Pennsylvania uses its estimates as only one of multiple measures in teacher evaluation. Second, rather than simply assigning the PVAAS estimate as a teacher's effectiveness rating, Pennsylvania instead uses the estimates to categorize a teacher's performance as "highly effective," "effective," or "ineffective." These two features of PVAAS implementation represent reasonable ways of dealing with VAA uncertainties.

The "cross-classified model"

The "cross-classified model" is similar to the layered model with a slight modification: it assumes that student achievement grows at a predictable and even rate over time.[58] The cross-classified model is unique in this respect. Other models allow for individual student growth[59] but do not explicitly model it as constant over time.

The difference between assuming constant growth and not assuming constant growth is subtle but important. Consider the following example that follows another nine students in a very small school. These students are in 5th grade; their scores on the state math assessment are as follows:

5th Grade Teacher	Student	4th grade Math Score	5th grade Math Score	Gain, 4th to 5th	Average Gain, 5th grade teacher	School Average Gain	Teacher Effect
1	1	39	47	8	8.00	3.33	4.67
	2	37	39	2			
	3	43	57	14			
2	4	35	39	4	-0.67	3.33	-4.00
	5	37	35	-2			
	6	39	35	-4			
3	7	45	51	6	2.67	3.33	-0.67
	8	47	45	-2			
	9	49	53	4			

If student growth was not assumed to be constant year-to-year (as with all VAA approaches except the cross-classified model), teacher effects are estimated based on the average of all student gains[60]. The calculation of effects of 5th-grade teachers would be the same as noted in the example above:

5th Grade Teacher	Student	4th grade Math Score	5th grade Math Score	Gain, 4th to 5th	Average Gain, 5th grade teacher	School Average Gain	Teacher Effect
1	1	39	47	8	8.00	3.33	4.67
	2	37	39	2			
	3	43	57	14			
2	4	35	39	4	-0.67	3.33	-4.00
	5	37	35	-2			
	6	39	35	-4			
3	7	45	51	6	2.67	3.33	-0.67
	8	47	45	-2			
	9	49	53	4			

The results are different if constant annual student growth is considered, as is the case with the cross-classified model. In such cases, gains associated with a particular teacher are adjusted by student average annual growth before teacher effects are estimated. A given teacher effect under this assumption is essentially the average of what is left over of annual student gains after each student's gain has been adjusted to take into account that student's constant annual growth.[61] The calculations under this assumption are as follows:

5th Grade Teacher	Student	4th grade Math Score	5th grade Math Score	Gain, 4th to 5th	Student Average Yearly Gain[i]	Student Adjusted Gain	Teacher Average Adjusted Gain	School Average Adjusted Gain	Teacher Effect
1	1	39	47	8	12	-4	-3.67	-0.22	-3.44
	2	37	39	2	7	-5			
	3	43	57	14	16	-2			
2	4	35	39	4	2	2	-0.33	-0.22	-0.11
	5	37	35	-2	-1	-1			
	6	39	35	-4	-2	-2			
3	7	45	51	6	1	5	3.33	-0.22	3.56
	8	47	45	-2	-4	2			
	9	49	53	4	1	3			

[i]Average yearly gain is typically estimated using data from several previous years

In this example, teacher effects differ substantially under the two different assumptions. In fact, the highest-scoring teacher under the assumption of non-constant growth (Teacher 1) is rated most poorly under the assumption of constant growth. Which one is right? This depends on your beliefs about the true nature of student growth and the contributions of teachers. The cross-classified model assumes that students would improve at some constant rate even without the added effects of teachers – so teachers should only be credited after that constant growth is taken into consideration. The other models make no such assumption.

In most other ways the cross-classified model is similar to the layered model. It considers data across multiple years, subjects and cohorts; it assumes teacher effects persist undiminished; and, it measures the effects of teachers across different years simultaneously, rather than piecemeal for each individual year. Unlike the layered model, the cross-classified model typically takes student and school characteristics into account.

The cross-classified model adjusts scores for (a) the constant growth of individual students, and (b) any student characteristics or previous educational experiences formally specified as part of the model. Once scores are adjusted, the difference between a specific teacher's average gain and the average gain of all teachers in the sample provides the teacher's effectiveness estimate.

Characteristics of the cross-classified model are summarized in Figure 4.

Figure 4.

Cross-classified Model

Outcomes modeled:

- Individual assessment scores across a multiple time periods (typically five years of assessment scores)
- Multiple subjects possible
- Multiple cohorts possible

Adjustments: Student and school characteristics

What differentiates it from other multiple-wave models

- Assumption that teacher effects persist undiminished
- Assumption that model can allow for predictable growth in student achievement

Key Strengths

- Inclusion of information from previous years
- Capability to include students with missing data (though missing scores must be estimated)
- Accommodation for constant student growth

Key Shortcomings

- Complexity; difficult to understand and explain
- Required specialized software
- Projected scores for students with missing data (may skew teacher effects)
- No consideration of point at which students start along the developmental scale
- Assumption of undiminished teacher effects (may be unreliable)
- Assumption of constant, predictable student growth (may be unreliable)
- Omission of statistical adjustments for student ability (may skew teacher effects)
- No accommodation for problems associated with using student assessment scores

Accountability programs based on the cross-classified model: VAA models based on the cross-classified model have been proposed, but to date they have been used primarily for research.[62]

The "persistence" model

A team led by RAND statistician Daniel McCaffrey created a multiple-wave model similar to the layered model but different in one major respect. In this

"persistence" model, teacher effects are not assumed to persist undiminished; rather, the model allows for the actual rate of persistence to be estimated.[63]

The persistence model shares several characteristics with the layered model. It can consider data across multiple years, subjects, and cohorts; it measures the effects of teachers simultaneously across different years; and it allows for individual student growth (although, unlike the cross-classified model, it does not model student growth as constant over time). In the persistence model, student scores are adjusted for (a) the growth unique to each student each year, and (b) any student characteristics or previous educational experiences formally specified as part of the model. Persistence of the effects of former teachers is always estimated as part of this model, and therefore its estimates always reflect this factor. As is true of some other models, the estimate of teacher effect in the persistence model is the difference between a specific teacher's average gain (based on adjusted scores) and the average gain of all teachers in the sample.

Characteristics of the persistence model are summarized in Figure 5.

Figure 5.

Persistence Model

Outcomes modeled:

- Individual assessment scores across a multiple time periods (typically five years of assessment scores)
- Multiple subjects possible
- Multiple cohorts possible

Adjustments: Student and school characteristics

What differentiates it from other multiple-wave models

- Estimated persistence of previous teacher effects (no assumption of undiminished persistence)
- No accommodation for student achievement growth

Key Strengths

- Inclusion of previous information
- Capability to include students with missing data (though missing scores must be estimated)
- Effects of previous teachers estimated, not assumed
- Assumption regarding constant student growth is not required

Key Shortcomings

- Complexity; difficulty to understand and explain
- Specialized software required
- Projected scores for students with missing data (may skew teacher effects)
- No consideration of point at which students start along the developmental scale
- Omission of statistical adjustments for student ability (may skew teacher effects)
- No accommodation for problems associated with using student assessment scores

Accountability programs based on the persistence model: VAA models based on the persistence model have been proposed, but to date they have been used primarily for research.[64]

Todd and Wolpin's "Cumulative within-child" mixed-effects model

A final model comes from the Educational Production Function literature. Todd and Wolpin[65] discuss several approaches to modeling educational outcomes. Their most

general model, of which all related models are special cases, is the "cumulative within-child" mixed-effects model. This model is similar to the other multiple-wave models described above in that it uses multiple sets of data. The key difference between the cumulative within-child model and other VAA models generally is that the cumulative within-child model explicitly includes consideration of unobserved characteristics of children that are permanently related to performance – labeled "child endowment" by Todd and Wolpin – as well as other historical and contemporary family and school characteristics.[66] The technical details of how this is done are beyond the scope of this guide,[67] but it is worth noting that cumulative-within-child model is distinguished by its attention to the specific ability of individual children.

This model considers child endowment in order to explicitly address a major shortcoming of other VAA models—the fact that it uses non-experimental data from situations in which teachers and students were not randomly grouped. Todd and Wolpin summarize the potential impact of omitting child endowment (as most VAA approaches do)[68]:

> …there is an implicit assumption of random assignment with respect to unobserved characteristics of children that are permanently related to performance (child endowment). If a particular teacher were systematically assigned to children with high endowments, the influence of the teacher on performance would be overstated. Averaging measured teacher gains over time, as is the practice in implementing TAAVS, will not eliminate this bias. However, it would seem possible to circumvent this problem by augmenting the specification to include child-specific fixed effects."

This model does provide the greatest accommodation for differences in child endowment when estimating teacher effects; however, modeling this construct can be highly problematic, as it is dependent on the validity of variables used to represent it.

Characteristics of the cumulative within-child model are summarized in Figure 6.

Figure 6.

Cumulative Within-Child Model

Outcomes modeled:

- Individual assessment scores across a multiple time periods (typically five years of assessment scores)
- Multiple subjects possible
- Multiple cohorts possible

Adjustments:

- Student and school characteristics
- Student past and present experiences
- Child-specific ability (modeled through statistical variables)

What differentiates it from other multiple-wave models

- Specific consideration of child's ability
- Flexibility to accommodate specific aspects of other VAA models

Key Strengths

- Inclusion of information from previous years
- Capability to include students with missing data (though missing scores must be estimated)
- Explicit allowance for non-random grouping of teachers and students through consideration of child's ability

Key Shortcomings

- Complexity; difficult to understand and explain
- Specialized software required
- Dependence on validity of variables used to estimate child endowment
- Projection of scores for students with missing data (may skew teacher effects)
- No consideration of point at which students start along the developmental scale
- No accommodation for problems associated with using student assessment scores

Accountability programs based on the cumulative within-student model: VAA models based on the cumulative within-student model have been used primarily for research.[69]

Additional VAA Models Currently in Development

Potential weaknesses of existing VAA models have led to the proposal of several "next-generation" approaches. Though several states have considered these alternatives, each has yet to be implemented as of this writing. Although they will not be discussed in detail here, the interested reader may consult Damian Betebenner's recent work on Markov Chain Models[70] and Rich Hill's work on Value Tables.[71]

Making Decisions Regarding VAA-based Accountability

This guide has been prepared primarily for the practitioner or policymaker needing to learn more about alternative models for VAA-based accountability. The detail provided above should make especially clear that the issues surrounding VAA-based programs are many and complex. As detailed in Table 2, approaches vary in their strengths, weaknesses, and suitability for any particular context. However, all VAA approaches share challenges that threaten the validity of teacher effect estimates they are designed to generate.

This section provides specific guidance towards choosing a VAA-based approach to become part of a district or state accountability program. The first half of this section covers issues specific to various VAA approaches to help guide the practitioner in crafting the optimum VAA-based approach for the targeted accountability program. No matter which model seems best suited to a particular context, however, general issues relating to the validity of VAA estimates must also be taken into consideration when determining how to incorporate VAA into systems of accountability. The second half of this section addresses challenges that threaten the validity of VAA teacher and school effect estimates *regardless of the VAA approach selected.*

Table 2. VAA Approaches: Strengths, Shortcomings, and Current Programs

		VAA Model						
		Single-wave models (two sets of measurements)			**Multiple-wave models** (multiple sets of measurements)			
		Gain Score	**Covariate Adjustment**		**Layered**	**Cross-Classified**	**Persistence**	**Cumulative Within-Child**
Key Strengths		• Simple; easy to understand • Easy to implement	• Simple; easy to understand • Easy to implement • Can model effects of prior educational experiences • Effects of previous teachers estimated, not assumed	•	• Data across >2 years • Can estimate scores for missing students • Requires no school/student information • Uses multiple years/subjects rather than student controls	• Data across >2 years • Can estimate scores for missing students Accommodates linear student growth	• Data across >2 years • Can estimate scores for missing students • Effects of previous teachers estimated, not assumed • Linear student growth assumption not required	• Data across >2 years • Can estimate scores for missing students • Adjusts for non-random grouping of teachers and students through adjustment for child's ability
Key Shortcomings		• No consideration of data from previous years • Students with missing data excluded • No consideration of point at which students start along developmental scale • Assumption of undiminished teacher effects • No statistical adjustments for student ability • No accommodation for problems of using student assessment scores	• No consideration of information from previous years • Students with missing data excluded • No consideration of point at which students start along developmental scale • No statistical adjustments for student ability • No accommodation for problems of using student assessment scores	•	• Complex; difficult to understand and explain • Specialized (proprietary) software required • Projected scores for students with missing data (may skew teacher effects) • No consideration of point at which students start along developmental scale • Assumption of undiminished teacher effects • No statistical adjustments for student ability • No accommodation for problems of using student assessment scores	• Complex; difficult to understand and explain • Specialized software required • Projected scores for students with missing data (may skew teacher effects) • No consideration of point at which students start along developmental scale • Assumption of undiminished teacher effects • Requires assumption of linear student growth • No statistical adjustments for student ability • No accommodation for problems of using student assessment scores	• Complex; difficulty to understand and explain • Specialized software required • Projected scores for students with missing data (may skew teacher effects) • No consideration of point at which students start along developmental scale • No statistical adjustments for student ability • No accommodation for problems of using student assessment scores	• Complex; difficult to understand and explain • Specialized software required • Dependence on validity of variables used to estimate child endowment • Projection of scores for students with missing data (may skew teacher effects) • No consideration of point at which students start along developmental scale • No accommodation for problems of using student assessment scores
Approach in Practice		Texas Growth Index	Dallas Value Added Assessment System (DVAAS)		Educational Value Added Assessment System (EVAAS)	None (used primarily for research)	None (used primarily for research)	None (used primarily for research

Selecting a VAA Approach

Multiple-wave Models: The Layered Model (EVAAS)

As noted above, the most common form of the layered model -- EVAAS -- is by far the most popular VAA-based approach with over 300 district and state programs in practice. Whereas most of the other approaches described above have been developed by researchers for research purposes, EVAAS is being aggressively marketed by a for-profit entity (The SAS Corporation) solely for accountability programs. As such, the overwhelming majority of practitioners faced with the prospect of implementing a VAA-based program as part of an educational accountability system are considering the EVAAS layered model.

As detailed above, EVAAS is distinguished primarily by two characteristics – its reliance on assessment scores across multiple years and multiple subjects rather than covariates for student and school characteristics, and its assumption that teacher/school effects persist undiminished as time passes. Research has suggested that using as many as five years of data across three separate assessments can render the use of statistical controls for student characteristics unnecessary;[72] as such, those building an EVAAS-based system are strongly advised to incorporate no fewer than than five years and three subjects worth of assessment data. Doing so doesn't guarantee that results will be unbiased in the absence of student statistical controls; additional studies have suggested that using even the expansive set of data suggested above fails to fully correct for differences in students not attributable to teachers or schools.[73]

A second distinguishing characteristic of EVAAS-based models is their assumption that teacher effects persist undiminished – that is, for example, the effect on a

student's achievement of having had a particular third-grade teacher will continue undiminished in all subsequent years. This turns out to be a problematic assumption; the most prominent work examining this assumption has suggested that teacher effects actually do diminish over time.[74] As a result, a particular teacher's VAA-based effect could end up unfairly higher or lower depending on whether she followed a relatively weaker or stronger teacher, as EVAAS would incorrectly credit previous teachers with effects that had actually diminished (and were therefore attributable to the more recent teacher). The only correction for this problem is to allow for diminishing teacher effects, a characteristic of the persistence model that has yet to be allowed by EVAAS.

To sum, although EVAAS is the most prominent VAA-based approach currently in use, it is not without its particular issues. One of these – the use of multiple scores rather than covariates – may be mitigated if a sufficient number of scores for each individual (such as scores across three subjects and five years) are used. The second – assuming undiminished teacher effects – is more pernicious and may very well result in unfairly crediting previous teachers for achievement associated with more recent teachers. This is a problem for which EVAAS currently allows no adjustment; as such, practitioners are strongly advised to either abandon EVAAS for more flexible models such as the persistence model, or – if this is not possible – to lower the stakes of EVAAS-based estimates and use them only as suggestive, for example, of where to target professional development resources (and not as a basis for personnel decisions).

Multiple-wave Models: The Cross-Classified Model

The cross-classified model's identifying characteristics are three-fold – (1) the assumption of constant (linear) annual growth for each student's achievement, (2) the

ability to use demographic covariates to adjust for student differences not attributable to teachers, and (3) the assumption that teacher effects persist undiminished. The assumption of linear achievement growth for each student (rather than unique year-to-year growth) may in fact be reasonable if an assessment's score scale is crafted to reflect an underlying construct that is truly believed to grow linearly. This is a decision that must be made in consultation with assessment developers; if it is in fact warranted, then use of the cross-classified model may be recommended. The cross-classified model can be used with data across multiple years or subjects; it can also handle statistical adjustments for student differences. If the use of data across multiple years and subjects is to take the place of student covariates (as is the case with EVAAS), individual scores across at least five years and three subjects are recommended. Finally, as was the case with the EVAAS above, the assumption of undiminished teacher effects is problematic and warrants that that assumption be relaxed in practice or that a different model (based not on that assumption) be used.

Multiple-wave Models: The Persistence and Cumulative Within-Child Models

The persistence model and cumulative within-child models are essentially the same with one distinction – the modeling of student differences via child inherited attributes, which is possible with the persistence model and compulsory with the cumulative within-child model. In practice, the unavailability of measures of child inherited attributes for large-scale accountability programs renders the cumulative within-child model less applicable for purposes outside of research. In fact, although the persistence model has yet to be used as part of a formal accountability program, its

flexibility situates it well for that use. It can handle data across multiple years and subjects (a strength of the EVAAS layered model) as well as covariates to adjust for student differences; it requires no assumption about persistence of teacher effects, allowing them to be estimated as part of the model estimation; also, although typically it does not do so, the persistence model could allow for linear growth (as does the cross-classified model). In most cases the persistence model provides for adjustments that relax assumptions of other models, so it is an ideal alternative among VAA-based approaches. It still does not overcome many problems of VAA-based approaches in general, which are discussed in the latter part of this section.

Single-wave Models: The Gain Score and Covariate Adjustment Models

In the infancy of the assessment-based accountability movement, single-wave models were often the only alternatives available due to the lack of more than two years of data. Today they are less prominent as many testing programs have been in place long enough to generate several years worth of achievement scores. As such, it is increasingly less likely that a practitioner would be called upon to consider implementation of a single-wave model. Nevertheless, it is helpful to have guidance on the the two primary single-wave models – the gain score model and the covariate adjustment model – and how these two models differ.

The primary difference between the gain score model and covariate adjustment model is the same as that which distinguishes the layered and persistence models – the assumption regarding persistence of teacher effects. The gain score model is similar to the layered model in that teacher effects are assumed to persist undiminished; on the other hand, the covariate adjustment model – like the persistence model – makes no such

assumption. As noted before, recent research has suggested a diminishment of teacher effects over time, rendering the covariate adjustment model preferable to the gain score model (similar to the persistence model's preference over the layered model). If the policymaker must choose between the two, it appears that the covariate adjustment model is preferable.

Issues Common to all VAA Approaches

The previous section provided guidance regarding selection among several alternative VAA-based approaches. However, no matter which model seems best suited to a particular context, general issues relating to the validity of VAA estimates must also be taken into consideration. Implementers are always wise to remember that because it is generally impossible to randomly group teachers and students, any VAA estimate of teacher effects may be influenced by other factors, as detailed above. While a particular approach may attempt to deal with this issue by using statistical adjustments, no set of adjustments can fully compensate for the lack of randomization. Furthermore, important questions remain about whether student achievement scores are appropriate measures of teacher effects; whether achievement tests are appropriately designed; whether and how assessment errors may affect estimates; and when assessments are best administered. Although these are crucial and complex questions, no VAA approach yet takes them into account.

No VAA approach yet provides perfectly valid estimates of teacher and school contributions to student learning. The tasks to which VAA approaches are applied simply don’t allow for a single correct or perfect way of accomplishing them. Trade-offs and risky assumptions are required in any case, so any given model is necessarily going

to be imperfect. This should not be taken as criticism toward the developers of any given approach; rather, it's meant as a criticism of the construct validity of any approach as applied as a tool of accountability.

Above all, expectations for what any VAA-based tool can reasonably accomplish should be tempered, and the use of its estimates should be judicious. As the introductory quote by George Box noted, all models are wrong—including models based on VAA approaches. The weaknesses of VAA detailed in this guide render VAA inadvisable as the cornerstone of a system of teacher or school accountability. Any teacher or school effect estimated from VAA models should be taken as only that—an estimate. VAA-based estimates may help identify teachers who appear to be successful as well as those who appear to need assistance in improving their practice. However, until more is known about the accuracy of and tradeoffs between VAA approaches, VAA-based estimates should never serve as a single indicator of teacher effectiveness, and high stakes decisions should never be made primarily on the basis of VAA-based estimates.

Notes and References

[1] Carey, K. (2004). "The real value of teachers." *Thinking K-16, 8*(1), 3-32.

[2] Sanders, W. L. (2005, June). *A summary of conclusions drawn from longitudinal analyses of student achievement data over the past 22 years (1982—2004)*. Presentation to Governors Education Symposium, Asheville, NC.

[3] Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer, and G. N. Wilkinson, (eds.) *Robustness in Statistics* (pp. 201-236). New York: Academic Press.

[4] See, for example, McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation. A second resource is a special edition of the Journal of Educational and Behavioral Statistics devoted to Value Added Assessment, Wainer, H. (Ed.). (2004). Value-Added Assessment [Special Issue]. *Journal of Educational and Behavioral Statistics, 29*(1).

[5] Braun, H. (2005, September). Using student progress to evaluate teachers: A primer on value-added models. [Policy Information Perspective]. New Jersey: ETS.

[6] No Child Left Behind Act of 2001. (2002). Public Law 107-110 Statute 1425.

[7] California Department of Education (n.d.). *Academic Performance Index (API)*. Retrieved January 6, 2006, from http://www.cde.ca.gov/ta/ac/ap/.

[8] Department of Accountability and Data Quality, Texas Education Agency (2005, June). 2005 accountability manual. Retrieved 1/10/06 from http://www.tea.state.tx.us/perfreport/account/2005/manual.

[9] The North Carolina system is based on changes in the average score of students matched students across two successive years. More detail can be found at *ABCs* (n.d.). Retrieved January 6, 2006, from http://abcs.ncpublicschools.org/abcs/.

[10] The Arizona system is based on the percentage of students maintaining or improving their standing relative to grade-specific performance stanines. More detail can be found at *Analysis of the Arizona Measure of Academic Progress, 2000-2001* (n.d.). Retrieved January 6, 2006, from http://ade.az.gov/Researchpolicy/academicprog/MAPsummarypiece.pdf.

[11] See Todd & Wolpin (2003) for a nice description of production function models used in educational contexts. [Todd, P.E., & Wolpin, K. I. (2003). "On the specification and estimation of the production function for cognitive achievement," *Economic Journal 113*(February), pp. 3-33.]`

[12] Lord, F.M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin*, *72*(5), 336–337; for a more recent treatment see Mans, E. (1998). "covariance adjustment versus gain scores—revisited." *Psychological Methods 3*(3), pp. 309-327. See also Thum, Y.M. (2003). Measuring progress towards a goal: estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods & Research 32*(2), pp. 153-207

[13] Bryk, A.S., & Raudenbush, S. W. (1992). Hierarchical linear models: Applications and data analysis methods. Newbury Park, CA: Sage Publications.

[14] L. Burstein (1980). "The analysis of multilevel data in educational research and evaluation", in D. C. Berliner (Ed.), *Review of Research in Education 8*, pp. 158-223. Washington, DC: American Educational Research Association, 1980.

[15] Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics 38*, pp. 963–974.

[16] No Child Left Behind Act of 2001. (2002). Public Law 107-110 Statute 1425.

[17] Olsen, L., & Hoff, D. (2005, November). U.S. to pilot new gauge of 'growth'. *EdWeek 25*(13), pp. 1,16.

[18] McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation.

[19] Braun, H. (2005, September). Using student progress to evaluate teachers: A primer on value-added models. [Policy Information Perspective]. New Jersey: ETS.

[20] Braun, H. (2005, September). Using student progress to evaluate teachers: A primer on value-added models. [Policy Information Perspective]. New Jersey: ETS.

[21] Braun, H. (2005, September). Using student progress to evaluate teachers: A primer on value-added models. [Policy Information Perspective]. New Jersey: ETS p. 8.

[22] Braun, H. (2005, September). Using student progress to evaluate teachers: A primer on value-added models. [Policy Information Perspective]. New Jersey: ETS.

[23] Todd, P.E., & Wolpin, K. I. (2003). "On the specification and estimation of the production function for cognitive achievement," *Economic Journal 113*(February), pp. 3-33.

[24] Braun, H. (2005, September). Using student progress to evaluate teachers: A primer on value-added models. [Policy Information Perspective]. New Jersey: ETS, p. 9

[25] Braun, H. (2005, September). Using student progress to evaluate teachers: A primer on value-added models. [Policy Information Perspective]. New Jersey: ETS., p. 9

[26] For a detailed discussion of precision of VAA-based estimates see McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation.

[27] Braun, H. (2005, September). Using student progress to evaluate teachers: A primer on value-added models. [Policy Information Perspective]. New Jersey: ETS, p. 10

[28] See pp. 81-87 of McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation.

[29] Braun, H. (2005, September). Using student progress to evaluate teachers: A primer on value-added models. [Policy Information Perspective]. New Jersey: ETS, pp. 13-14 ; McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation (pp. 87-105).

[30] McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation (pp. 104-105)

[31] Todd, P.E., & Wolpin, K. I. (2003). "On the specification and estimation of the production function for cognitive achievement," *Economic Journal 113*(February), pp. 3-33; McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation.

[32] McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation; McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1).

[33] Todd, P.E., & Wolpin, K. I. (2003). "On the specification and estimation of the production function for cognitive achievement," *Economic Journal 113*(February), pp. 3-33.

[34] Northwest Evaluation Association has also proposed an approach based on the gain score model. See, for example, McCall, M. S., Kingsbury, G., G., & Olson, A (2004, April). Individual growth and school success. Technical Report. Lake Oswego, OR: NWEA.

[35] Department of Accountability and Data Quality, Texas Education Agency (2005, June). 2005 accountability manual. Retrieved 1/10/06 from http://www.tea.state.tx.us/perfreport/account/2005/manual.

[36] McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation, directs the interested reader to two examples of covariate adjustment models used in the EPF literature: (1) Hanushek, E. (1972). Education and race. Lexington, MA: D.C. Heath and Company.; and (2) Murnane, R.J. (1975). The impact of school resources on the learning of inner city children. Cambridge, MA: Ballinger Publishing Co.

[37] Webster, W. & Mendro, R. (1997). The Dallas value-added accountability system, in J. Millman (Ed.), *Grading teachers, grading schools: is student achievement a valid evaluation measure?* (pp. 81-99). Thousand Oaks, CA: Corwin Press, Inc.

[38] McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation.

[39] Todd, P.E., & Wolpin, K. I. (2003). "On the specification and estimation of the production function for cognitive achievement," *Economic Journal 113*(February), pp. 3-33.

[40] For details about EVAAS see *Schooling Effectiveness, SAS® EVAAS® for K-12* (n.d.). Retrieved January 4, 2006, from http://www.sas.com/govedu/edu/services/effectiveness.html.

[41] Ballou, D., Sanders, W., & Wright, P. (2003). Controlling for students' background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), p. 60.

[42] *Schooling Effectiveness, SAS® EVAAS® for K-12* (n.d.). Retrieved January 4, 2006, from http://www.sas.com/govedu/edu/services/effectiveness.html.

[43] See, for example, Todd, P.E., & Wolpin, K. I. (2003). "On the specification and estimation of the production function for cognitive achievement," *Economic Journal 113*(February), pp. 3-33; Wainer, H. (Ed.). (2004). Value-Added Assessment [Special Issue]. *Journal of Educational and Behavioral Statistics, 29*(1).

[44] Braun, H. (2005, September). Using student progress to evaluate teachers: A primer on value-added models. [Policy Information Perspective]. New Jersey: ETS, provides a nontechnical description of the EVAAS. As he notes, the best technical description of EVAAS can be found in Sanders, W. L., Saxton, A., & Horn, B. (1997). The Tennessee Value-Added Assessment System: a quantitative outcomes-based approach to educational assessment," in J. Millman (Ed.), *Grading teachers, grading schools: is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press, Inc. Other descriptions can be found in Helland, K. (2002). Value added assessment – a school directors' handbook." Olympia, WA: Evergreen Freedom Foundation; and Ballou, D., Sanders, W., & Wright, P. (2003). Controlling for students' background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), p. 60.

[45] Ballou, D., Sanders, W., & Wright, P. (2003). Controlling for students' background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), p. 60.

[46] Todd, P.E., & Wolpin, K. I. (2003). "On the specification and estimation of the production function for cognitive achievement," *Economic Journal 113*(February), pp. 3-33; McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1)

[47] McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1)

[48] The Center for Greater Philadelphia (n.d.). *Value-added assessment across the nation*. Retrieved January 6, 2006, from http://www.cgp.upenn.edu/ope_nation.html

[49] The Center for Greater Philadelphia (n.d.). *Value-added assessment in Colorado*. Retrieved January 6, 2006, from http://www.cgp.upenn.edu/ope_co.html.

[50] Seattle Public Schools (n.d.). *Understanding value-added data in Seattle.* Retrieved January 6, 2006, from http://www.seattleschools.org/area/valueadded/vahelp.xml; The Center for Greater Philadelphia (n.d.). *Value-added assessment in Washington*. Retrieved January 6, 2006, from http://www.cgp.upenn.edu/ope_wa.html.

[51] Minneapolis Public Schools (n.d.). *Accountability*. Retrieved January 6, 2006, from http://www.mpls.k12.mn.us/Accountability.html

[52] Iowa Association of School Boards (n.d.). *Introducing Iowa Value-Added Assessment System (IVAAS).* Retrieved January 6, 2006, from http://www.ia-sb.org/services/ivaas.pdf

[53] Butry, B. M. (2005). Value-added pilot project begins this fall. *On Board Online 6*(14). Retrieved January 7, 2006, from http://www.nyssba.org/ScriptContent/VA_Custom/va_cm/contentpagedisplay.cfm?Content_ID=3971&SearchWord=value-added%20assessment.

[54] Battelle for Kids (n.d.). *BFK: Project SOAR.* Retrieved January 6, 2006, from http://www.battelleforkids.org/b4k/rt/about/our_work/improve/SOAR

[55] Battelle for Kids (n.d.). *Value-added analysis:a critical component of Ohio's accountability system.* Retrieved January 6, 2006, from http://www.battelleforkids.com/b4k/rt/null?exclusive=filemgr.download&file_id=4521&rtcontentdisposition=filename%3DVAWhitepaper.pdf

[56] University of Dayton Alumni News, "What makes a great teacher?" (2004, June). Retrieved January 7, 2006, from http://alumni.udayton.edu/np_story.asp?storyID=1624.

[57] The Center for Greater Philadelphia (n.d.). *A New System of Accountability*. Retrieved January 6, 2006, from http://www.cgp.upenn.edu/ope_new_system.html

[58] The cross-classified model could incorporate any parametric growth model, though linear growth is most often assumed in practice.

[59] Other models typically allow for idiosyncratic growth through residual terms specific to student-year combinations

[60] In this case, differences between the any particular student's gain and the average gain represent the effect that is idiosyncratic to that student in that year.

[61] In practice the calculations are more complex, including statistical adjustments for student and school characteristics, but the general idea remains consistent with this example.

[62] Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools, *Teachers College Record 104*, 1525-1567.

[63] McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation; McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1).

[64] McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation; McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1).

[65] Todd, P.E., & Wolpin, K. I. (2003). "On the specification and estimation of the production function for cognitive achievement," *Economic Journal 113*(February), pp. 3-33.

[66] Because "child endowment" is measured with error, the cumulative within-child approach applies an instrumental variables technique (in which lagged values of student and siblings serve as instruments) to a model that includes within-child fixed effects to model it. See Todd, P.E., & Wolpin, K. I. (2003). "On the specification and estimation of the production function for cognitive achievement," *Economic Journal 113*(February), pp. 3-33 for more detail.

[67] See Todd, P.E., & Wolpin, K. I. (2003). "On the specification and estimation of the production function for cognitive achievement," *Economic Journal 113*(February), pp. 3-33, for detail on this and other VAA approaches.

[68] Todd, P.E., & Wolpin, K. I. (2003). "On the specification and estimation of the production function for cognitive achievement," *Economic Journal 113*(February), pp. 3-33.

[69] Todd, P.E., & Wolpin, K. I. (2003). "On the specification and estimation of the production function for cognitive achievement," *Economic Journal 113*(February), pp. 3-33.

[70] Damian W. Betebenner (2005, June). *Performance standards in measures of educational effectiveness.* Paper presented at Large Scale Assessment Conference, San Antonio, TX.

[71] Hill, R., Gong, B., Marion, S., DePascale, C., Dunn, J., & Simpson, M. A. (2005, November). *Using value tables to explicitly value growth.* Presentation given at MARCES Conference on Longitudinal Modeling of Student Achievement. Retrieved January 10, 2006 from http://www.nciea.org/publications/MARCES_RHBGSMCDJDMS05.pdf.

[70] Todd, P.E., & Wolpin, K. I. (2003). "On the specification and estimation of the production function for cognitive achievement," *Economic Journal 113*(February), pp. 3-33.

[71] McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation; McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1).

[72] Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for students' background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), p. 37-66.

[73] Raudenbush, Stephen W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*(1), p. 121-130.

[74] McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (Report MG-158). Santa Monica, CA: RAND Corporation.