COMMISSIONED WHITE PAPER

Modeling Growth for Accountability and Program Evaluation: An Introduction for Wisconsin Educators

July 2009

Bob Lissitz
University of Maryland

and

Harold Doran
American Institutes for Research

This paper was commissioned by the Wisconsin Department of Public Instruction to promote greater understanding of growth and value-added models of student achievement, particularly as they might be applied within Wisconsin.

TABLE OF CONTENTS

1. Introduction

A defining trend in the world of education has been the increasing appetite for data at all levels of our educational system. There is a growing demand for more and better data to be used for simple to complex analyses aimed at accountability, program evaluation, and school improvement. In recent years, data-driven decision making has become more than just a phrase; it has become a mantra for many school systems across the nation.

This is partly a result of state and federal accountability mandates, which depend on valid and reliable assessment data. But it goes beyond accountability. In *Balanced Scorecards and Management Data* Hess and Fullerton (2009) argue that collecting and analyzing data from a wide variety of sources have become requirements for educators who want to be successful. As the 2007 report *Managing for Results in America's Great City Schools* showed, the use of good management data has gained acceptance in schools across the country.

Good management data allows users to analyze results. For example, comparing results from a middle school's approach to math remediation against outcomes from other middle school programs can be very revealing. Comparative analyses like this can involve the whole spectrum of school activity. Data management in today's school systems includes benchmarking student performance, comparing the success rates of various intervention programs, and even analyses of financial systems. The balanced scorecard approach to management data is an illustration of something many educators at local, state, and national levels have already adopted.

Education has often been oriented within the humanities rather than the sciences. However, the movement of the field is clearly toward a more scientific orientation—greater emphasis on data, data analysis, experimentation, and systematic exploration of research questions. The federal government recently created the Institute for Education Sciences (IES), which places a heavy emphasis on experimental studies, a particularly difficult approach to education research. The new federal effort to define "What Works?" in K-12 education is another effort to require careful study and evidence-based claims before declaring an intervention or program is effective. The hope is that using evidence-based analyses will lead to sound conclusions and clear direction on how best to educate our children.

Growth models have generated a lot of excitement with their promise of increased support for data-driven decision making. In a nutshell, growth modeling is a statistical technique that analyzes the amount of change in a student's performance over time. Different growth models may answer slightly different questions about growth. In this paper we will explore different types of growth models and the questions they can best answer.

Federal accountability provisions under the No Child Left Behind (NCLB) Act of 2001 require all states to report annually the proportion proficient and above in reading and mathematics for multiple student subgroups in grades 3-8 and once in high school. NCLB

accountability provides a "snapshot" of student performance at one moment in time. This way of looking at school effectiveness is often referred to as a *status model*.

Status models serve an accountability purpose, but they cannot answer all the questions educators have regarding their students' performance. As snapshots of student performance, they do not answer questions about student growth. Status models cannot tell a teacher how much her students grew between September and May. Status models cannot tell parents whether their child grew as much in 7th grade as in 6th grade.

Yen (2007) published a list of common questions asked by parents, teachers and administrators. These questions reflect the understanding that education is a continuing process, and snapshots do not suffice in answering such questions (some questions have been reworded slightly for readability):

Parent Questions:
Did my child make a year's worth of progress this year?
Is my child growing enough to meet state standards?
Is my child growing as much in Math as Reading?
Did my child grow as much this year as last year?

Teacher Questions:
Did my students make a year's worth of progress this year?
Are my students growing enough to meet state standards?
Are there students with unusually low growth who need special attention?
Are there students with high growth who are nearing proficiency?

Administrator Questions:
Did the students in our school/district make a year's worth of progress?
Did our students grow in all content areas?
Did our students grow similarly across grades?
Does this program/intervention show as much growth as the other?

These questions concern the growth of our students. Knowing which students are growing well and which ones are not meeting growth expectations can be actionable information for a school or district. This is where *growth models* can provide information that status models cannot.

The methods for modeling growth vary, and often reflect the different methodological groundings of the analyst—whether an econometrician, a psychometrician, or an education researcher. These methodological underpinnings will be discussed in this paper only to the extent that they help to explain the general characteristics of a given model[1]

[1] Those readers interested in more technical descriptions can find excellent manuscripts prepared by the RAND Corporation (2001), the Spring 2004 edition of the *Journal of Educational and Behavioral Statistics*, or other papers published by statisticians on the topic (Lissitz; 2005, Lissitz, 2006; Lockwood, Doran, McCaffrey, 2001; Lockwood, McCaffrey, Mariano, Setodji, 2007).

When considering using one of the models described in this paper, readers will need to think about whether their agency is currently collecting data that could be analyzed within the model. For example, if a superintendent wants to know which reading intervention program performed best, the district would have to be systematically tracking which interventions children are exposed to within the district.

Growth modeling takes many forms. Models of interest in one context—for example in a small district—may be of less interest in another. Different school districts can and should try different approaches to questions regarding growth. Similarly, the Wisconsin Department of Public Instruction (DPI) might want to approach growth models in a different way than districts. Certainly, some models are simpler than others. This paper will offer some recommendations and considerations, both from a technical and policy perspective. First, let's examine the various approaches to growth modeling.

2. Modeling Growth for Accountability

In this section we review the use of growth models for making adequate yearly progress decisions for NCLB accountability. States are required to apply to the U.S. Department of Education (USED) for permission to use a growth model in their accountability plans. As of this writing, 14 states have approval. Two kinds of growth model have been approved: *growth-towards-the-standards models* and *value tables*. As a result of limitations imposed by USED, value-added models that control for student characteristics are not currently approvable for federal accountability.[2] These more complex models, which nonetheless show promise for research and program evaluation, are therefore described later in this paper.

This section discusses how different growth models are used for NCLB accountability. The Council of Chief State School Officers (CCSSO) *Implementer's Guide to Growth* (2007) provides a strong foundation for this discussion, and their more recent document (2009) summarizes the federal growth model pilot program from 2005 to 2008. Our aim is to extend that discussion by considering more recent developments in growth models as implemented by states as well as emerging research relevant to growth models.

2.1 Accountability under No Child Left Behind

While there is no consensus in the academic literature on the best use of test score data for comparing school performance, federal law under NCLB codified one methodology commonly known as the status model—a method that computes the percentage of

[2] It is unclear at this time whether USED, under the Obama administration and Secretary of Education Arne Duncan, will maintain these limitations (see section 2.2). The upcoming reauthorization of NCLB may also include new flexibility or requirements for using growth models for accountability.

students at or above proficient[3] from a single test instance and compares this to an expected threshold (known as the annual measurable objective or AMO). This method has well-known limitations (Meyer, 1997), and is often criticized on the basis that it evaluates schools only on a single set of test scores. While the status model may be a good starting point, growth modeling and multiple measures across time are needed to answer important questions that can't be answered by a status model alone.

The status model was historically a popular choice as longitudinal test data were not commonly available. However, NCLB introduced testing requirements whereby individual students are tested in reading and mathematics at least once per year in grades 3 through 8 and once in high school. These annual testing requirements represent a substantial jump in the amount of longitudinal data educators can use and form a cornerstone for measuring student progress over time.

2.2 NCLB Growth Model Pilot Program

In 2005, the U.S. Secretary of Education announced a pilot program allowing states to implement a growth model in addition to the required status model to make accountability decisions. The pilot program had a maximum of 10 states that could be approved. To be approved, states were required to have a fully approved state assessment system and to obtain approval for their proposed growth model from a peer review process.

A fully approved assessment system refers to states having submitted sufficient evidence regarding the critical elements enumerated in the peer review guidance (USED, 2007) for all assessments used for NCLB-related purposes. The second criterion, approval from the peer reviewers, was guided by the "bright line principles" – the core principles of NCLB that could not be compromised – espoused by Secretary Spellings (2005). A group of peer reviewers, consisting of experts in educational measurement, policy, and practice, were asked by the USED to review each of the submitted state proposals. Approval was based on whether the proposed models met criteria such as these:

- Set annual growth targets that lead to 100% proficiency by 2013-14
- Include all students and student subgroups
- Account for Reading/Language Arts and Mathematics separately
- Do not base growth targets on individual student background characteristics

These criteria are still in use, although the initial 10-state restriction to the pilot program has been lifted. More substantive details on the criteria can be found at the USED web site at http://www.ed.gov/admins/lead/account/growthmodel/index.html.

[3] Each state defines what is meant by proficiency relative to the knowledge and skills defined in its own academic standards. Test scores required to attain a "proficient" performance level are determined through a standard-setting process involving empirical data, educator judgment, and policy considerations.

2.3 Growth Models Approved for NCLB Accountability

The following sections provide examples of models that have been approved for NCLB accountability. We do not include exhaustive descriptions of each state model. For additional details of the implementation of each state growth model, see the USED website (above) or the CCSSO papers (2007, 2009).

2.3.1 Growth-Towards-the-Standards Models

Growth models being used for federal accountability must somehow incorporate proficiency into their calculations. To this end, a class of growth models referred to as growth-towards-the-standards models (GTS) have been proposed (Doran & Izumi, 2004; Sanders, 2006; Thum, 2001). These models relate a student growth trajectory (test scores over time) to a proficiency cut score (the test score that separates proficient from not proficient) on the state assessment.

GTS models tend to be concerned with students who are currently scoring below the proficient cut score, but their growth trajectory suggests that they will be proficient within a finite number of years. In many respects, GTS models are designed to answer the question: "Given the observed student growth from time 1 to time 2, is the student on-track to reach proficiency by time 3?"

There are two GTS variations: the yearly target and future projection models. Each type is discussed below.

Yearly-Target Models

The GTS yearly-target models evaluate students on whether they have met specific yearly targets in each grade. These targets are based on growth trajectories for individual students, beginning with their initial (observed) score and ending with a proficient score in a later grade. The table below shows states using this type of model, the year the model was approved, the number of years within which students must become proficient (trajectory), and whether the model requires a vertical scale.[4]

[4] A vertical scale uses a single scale across grade levels. This allows a child's performance on a test to be compared from one grade to the next on the same scale, making it very easy to calculate growth based on the scale. A vertical scale is illustrated in figure 4, and vertical scales are discussed in more detail later in this section and again in section 3.2.3.

Table 1. Summary of approved state growth models: growth towards the standard with yearly targets

State	Year Approved	Trajectory (years)	Vertical Scale
Alaska	2007	4[a]	No
Arizona	2007	3	Yes
Arkansas	2006	Variable[b]	Yes
Colorado	2009	3	Yes
Florida	2007	3	Yes
Iowa	2007	4	Yes
Missouri	2008	4[c]	Yes
North Carolina	2006	3	Yes

a. If a student is in grades 3-6, then four years. Otherwise, the number of years is grade 10 minus the current grade.
b. By grade 8.
c. Within four years or by grade 8, whichever comes first.

Let's take a look at how this kind of model works. Assume we have a group of students tested for the first time in grade 3. For purposes of growth, we are interested in determining whether the student is on track to reach proficiency by grade 7. One option states have pursued is to use the first score as the baseline and to construct a growth trajectory extending from the base score to the proficient cut score in grade 7. When this model is used, a student is deemed to have made Adequate Yearly Progress (AYP) if their observed scores in grade 4, 5, or 6 are equal to or higher than the yearly growth targets that indicate the student is going to reach the proficiency cut score in grade 7.

For example, assume the state measures progress on a vertical scale ranging from 100 to 500 and that the cut score for proficiency in grade 7 is 350. Assume student A has a grade 3 scaled score of 160. It is now possible to construct a series of intermediate targets for grade 4, 5, and 6 using only the grade 3 scale score and the grade 7 proficient cut score.

The yearly change in scale score required to reach proficiency by grade 7 can be found by dividing the difference in scores by the number of years to reach grade 7:

$$\frac{350-160}{7-3}=48$$

This method indicates that this student must improve by at least 48 scaled score points per year to be on track to reach the grade 7 cut score for proficiency. If the student's scaled scores in grade 4, 5, or 6 are equal to or greater than the yearly targets, the student is included in the AYP calculations as having made AYP. Table 2 shows what the intermediate targets would be expressed in scale score units.

Table 2. Sample intermediate targets for a GTS yearly-target model.

Grade 3	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**
Observed Score	Target 1	Target 2	Target 3	Target 4
160	206	254	302	350

In this example, there are intermediate targets for grades 4, 5, and 6. Grade 7 represents the end of the timeline, at which point the goal should be met. In this grade, the student must reach a scale score of 350 (the grade 7 cut score) in order to be included in accountability calculations as being proficient.

Figure 1 provides a visual interpretation of this example. The black dot in grade 3 represents the observed scale score for this student. From this baseline score, a line extends to the grade 7 proficient cut score. In order to make growth for AYP, this student's scores in grades 4, 5, and 6 must be equal to or higher than the dark line. Notice that the student is expected to be making regular improvement from year to year, finally reaching the proficiency level at year 7.

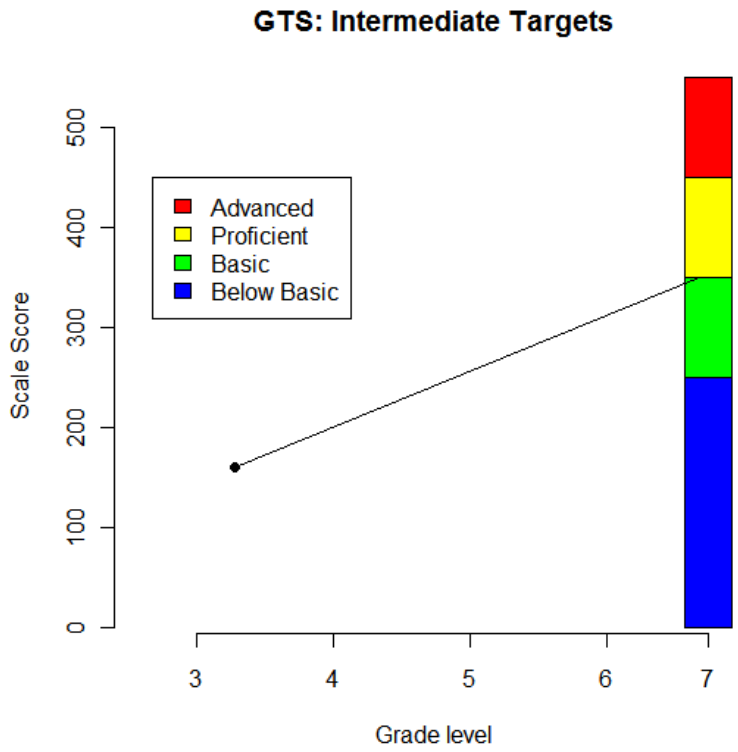

Figure 1. Intermediate targets for a GTS yearly-target model.

Projection Models

A second variation of the GTS model is one that uses student growth to determine if the student is likely to be proficient in the future by making a statistical projection. This GTS projection model only differs slightly from the GTS yearly-target model above. In the yearly-target model, all that is needed is a baseline score and the cut score at the end of the timeline. In the Projection model, at least three scores are required: at least two scores to measure growth, and the cut score for proficiency at the end of the timeline. The table below shows states that use this type of model, the year the model was approved, the

years allowed to become proficient (trajectory), and whether the model requires a vertical scale.

Table 3. Summary of approved state growth models: growth towards the standards with projected proficiency

State	Year Approved	Trajectory (years)	Vertical Scale
Ohio	2007	Variable[a]	No
Pennsylvania	2009	2	No
Tennessee	2006	3	No
Texas	2009	Variable[b]	No

a. One year beyond a school's grade configuration. For example, for a K-5 school, projections are made to grade 6.
b. Projected to be proficient by the next "high stakes" grade: 5, 8, or 11.

Extending the yearly-target example, assume student A has now been tested in grades 3 and 4 and has scores of 160 and 220, respectively. Further assume we are still interested in projecting to grade 7 as before.

In the simplest case, growth can be measured by subtracting the grade 3 scaled score from the grade 4 scaled score, yielding a growth estimate of (220-160) = 60. Now, assuming that this student continues to grow at the rate of 60 points per year, we can construct the grade 7 future projection as: 220 + (60 *x* 3) = 400. Recall that the cut score for proficiency in grade 7 is 350. , Since 400 > 350, we would claim that this student is "on track," or projected to be proficient in grade 7.

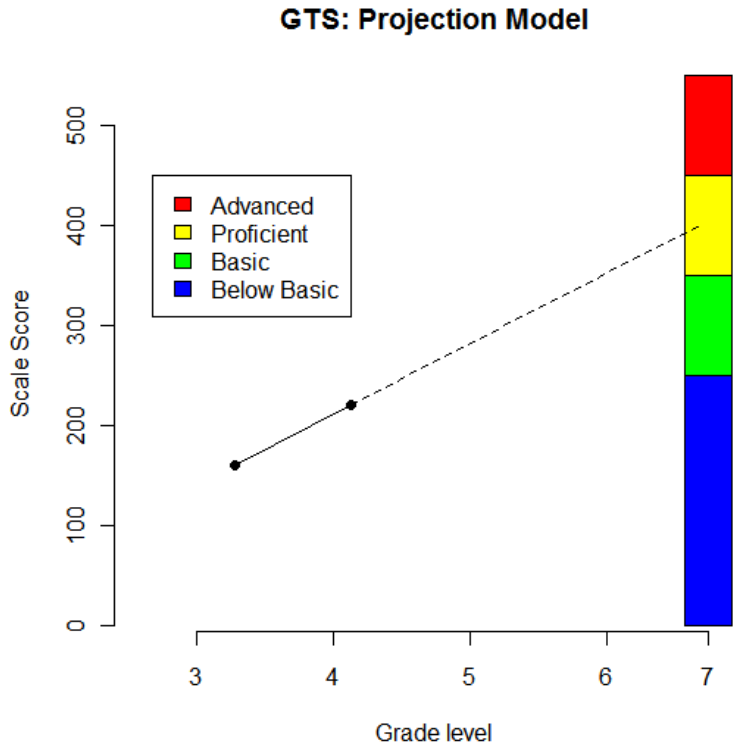

Figure 2. Projected growth in a GTS projection model

Figure 2 illustrates this model. In this figure, the two dark dots are the observed scores collected in grades 3 and 4. The dark line connecting them is the growth slope from grade 3 to 4. The dashed line is the linear extrapolation extending this growth to grade 7.

Technical Considerations for Growth-Towards-the-Standards Models

In both the yearly target and projection models, the growth estimates (trajectories) are formed from the single vertical score scale on the state assessment. A vertical scale makes the computation of growth straightforward. Scores at different grades can be directly compared because they are, by definition, on the same scale. However, many states lack a vertical scale and still desire a GTS model similar to the ones presented in this section.

Some experts believe there is nothing that precludes a state from using one of the two prior frameworks in the absence of a vertical scale. In fact, the models used in Ohio and Tennessee are GTS projection models and do not make use of the state vertical scale (Wright, Sanders & Rivers, 2006). Methods that do not rely on a vertical scale include standardizing scale scores and converting student scores to percentiles. Schafer (2006) proposes an approach that only requires stability in the scale from year to year (in other words horizontal equating).

In the second example above, growth was measured by taking the difference between the grade 3 and 4 scaled scores. This is indeed a raw growth rate and it is simply determined. However, it may not be the most robust method for determining student-level growth, especially if more than two scores are used. Student-level growth or projections could be estimated using mixed-effects linear models (Doran & Lockwood, 2004), or covariate adjustment approaches, for example.

The GTS projection model treats the projected score as if it were fixed, or known with certainty. From a statistical perspective, this is unwise since these projected scores cannot be predicted with certainty. In other words, we cannot be certain that a student projected to be proficient actually will be proficient. We only can state with some probability that a student is likely to be proficient.

One option is to assume the projected scores arise from a probability distribution and that the future proficiency status of an individual can only be expressed probabilistically. Figure 3 provides one illustration of how this may be accomplished. Rather than treating the future projection as fixed, we might assume the projected score is normally distributed. Subsequently, the goal is to find the portion of the probability distribution that falls above the proficient score (the yellow shaded area). This portion yields the probability that the student will be proficient in grade 7.

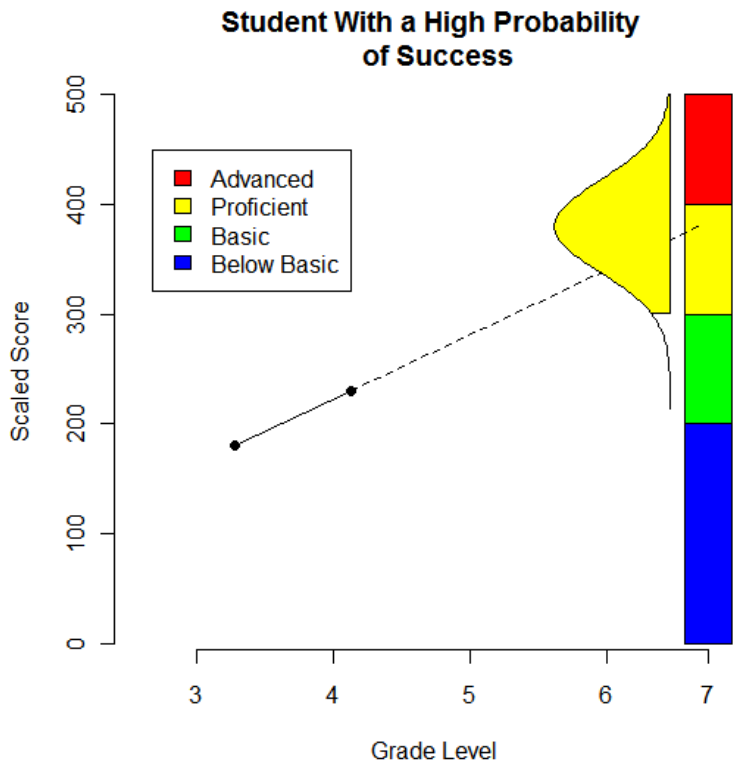

Figure 3. Projected proficiency as a probability distribution

There are other methods for forming the conditional probabilities or for setting conditional expectations. For instance, Betebenner and Shang (2007) argue that the single expectation of universal proficiency in GTS models is unreasonable. Instead, they demonstrate how the future status of individuals should take into consideration prior achievement of similar individuals and set reasonable growth targets based on the existence of students who have met them (Linn, 2003). Linn has discussed the idea that what constitutes expected progress should be reasonable and in part determined by what prior data patterns have shown to be possible. While this idea is important, it is currently in contrast to what the USED growth model pilot program demands, which is that students be held to the proficiency standard as the desired target of academic achievement, regardless of whether that is realistic. We will say a little more about this approach in section 3.1.2 on growth percentiles.

Last, many of the state GTS models proposed for the federal pilot program only implement the model for students currently scoring below the proficient cut score in their current grade. However, some students who score above proficient may be on a downwards trajectory such that they are unlikely to be proficient in the future. Most state growth models avoid these cases as they would negatively affect the AYP calculations.

Ignoring these downward moving students can be criticized on many levels. First, it is clearly designed to only add favor to a school in terms of making AYP. Second, there is

almost always a negative correlation between initial status and growth. That is, students scoring low in the beginning tend to grow at faster rates than students that are initially scoring high. This is called regression to the mean and is generally considered to be a statistical artifact – that is, not a function of effective teaching. In other words, GTS models may overestimate achievement to the extent that low scoring students appear to be on track only as a function of the regression error and not due to instructional quality.

2.3.2 Value Tables

A value-table approach is another method that states have chosen for measuring student growth over time. They have assumed the name *value table* because a series of points are awarded to individuals based on their growth from a low performance category to a higher performance category based on how much that change is valued. In other words, larger growth is valued more than smaller growth and larger growth would therefore receive more points. The table below shows states that use this type of model, the year the model was approved, and whether the model requires a vertical scale.

Table 4. Summary of approved state growth models: value tables

State	Year Approved	Vertical Scale
Delaware	2006	No
Michigan	2009	No
Minnesota	2009	No

There are multiple variations on the value-table approach (Betebenner, 2005; Hill, Marion, DePascale, Dunn, Simpson, 2006; Martineau, 2007). Some of these approaches, such as the transition matrix approach proposed by Betebenner (2005), use more formal statistical methods to monitor the transitions of students between performance categories over time. Other approaches use the observed scores of students and only credit schools for students that move between performance categories.

Table 5 shows the value-table approach taken from the Delaware state assessment system. In Delaware, levels below proficiency (i.e., Basic and Below Basic) are further subdivided into multiple performance levels to make more granular growth estimates. The Basic category is subdivided into Level 2A and Level 2B. The Below Basic category is subdivided into Level 1A and 1B. This allows for a student to move from the lower end of Basic (Level 2A) to the higher end of Basic (Level 2B) and still allow the school to receive some credit for that growth even though the growth occurs within the same performance category and the student is not yet Proficient.

In the Delaware value-table, points are only awarded to schools when students move across a performance category (or performance category subdivision) that moves them closer to proficiency than they were in year 1. For instance, 175 points are awarded to students that move from Level 1B in year 1 to Level 2A in year 2. However, 0 points are awarded to students that move from Level 2B in year 1 to Level 2A in year 2.

Table 5. The Delaware value-table model.

Year 1 Level	Year 2 Level				
	Level 1A	Level 1B	Level 2A	Level 2B	Proficient
Level 1A	0	150	225	250	300
Level 1B	0	0	175	225	300
Level 2A	0	0	0	200	300
Level 2B	0	0	0	0	300
Proficient	0	0	0	0	300

Clearly, this model does not use or require a vertical scale. All that is required are performance categories for each grade and a set of value points for each cell in the value table. However, it does assume some articulation of the standards from one grade to the next. That is, we assume that students with scores in higher performance categories in year 2 have improved in their knowledge and skills when compared to the prior year performance.

Technical Considerations for Value-Table Models

There are at least three technical issues for consideration with value-table approaches. The first is how to derive the points assigned for student transitions. Test scores or current test development procedures do not carry any information that can be used to derive these scores empirically. As such, states have used human judgment to determine the value scores. This is seen by some as a favorable practice since a public discussion of the implications for moving students from Below Basic to Basic or from Basic to Proficient can be an important matter with resource implications. One possible outcome is that a school could be rewarded for moving students from Proficient to Advanced. NCLB is often interpreted in a way that does not value such an effort, but instead encourages that resources be placed at moving the students to proficiency.

Second, if performance categories are to be subdivided into multiple categories, how can these subdivisions be made in a reasonable and defensible manner? Martineau (2007) has proposed that subdivisions can be made on the basis of the standard error of measurement on the test scale such that changes from one level to the next level must be larger than what would be observed from measurement error alone. Again, judgment may also be used.

Last, using these scores to make AYP decisions can be somewhat of a challenge. In Delaware, the average number of points earned by a school is compared to an annual measurable objective (AMO). A school makes AYP if their average points earned from the value-table are equal to or greater than the AMO.

Clearly, such a computation differs from the traditional methods used for computing status as the status AMOs are based on the percentage of students scoring at or above proficient. NCLB currently does not permit a state to have multiple AMOs for the same subject; therefore, negotiating this method of AYP computation with the USED would be

a likely challenge. Of course, a state or local school system can still adopt this approach for its own purposes and not make an effort to receive government approval.

2.4 Summary of Models for NCLB Accountability

A common thread in the different approaches states have taken to modeling growth for NCLB accountability is a focus on comparing the academic growth of individual students as measured by current state assessments to some target or objective for accountability purposes. The GTS yearly-target models use one baseline test score to create intermediate targets, against which students are evaluated. The GTS Projection models use two or more test scores to create a projected score for each student, which is then compared against the test score required for proficiency. The value-table approaches award points to students for moving up in proficiency categories, and the number of points earned is then compared against an objective.

3. Modeling Growth for Research and Program Evaluation

Growth modeling has many applications outside NCLB accountability. In light of the restrictions imposed by USED on NCLB growth models, the most interesting applications and most complex models are more promising for purposes such as research and program evaluation. In particular, value-added models that control for student background characteristics or other covariates are generating a great deal of interest in education. In this section we first consider simple normative growth models, and then discuss the more complex value-added models.

3.1 Normative Growth Models

Perhaps the most basic and intuitively appealing way of modeling growth is simply to subtract last year's test score from this year's test score. This difference in scores is the magnitude of growth. For example, if a student obtained a scale score of 430 on the third grade WKCE mathematics test and a scale score of 440 on the fourth grade test, the student has grown by 10 scale score points (440 – 430) in one year.

This way of modeling growth has among its advantages that it is easy to calculate, report, and understand. A significant disadvantage, however, is that additional information is required to meaningfully interpret that growth. Simply reporting the magnitude of growth does not supply enough information.

One reason is that even well-designed vertical scales in education "are, at best, quasi-interval" (Betebenner, 2008). In other words, the same change in scale score points might represent different amounts of learning depending on where a student is on the scale. That is unlike the scale on a yardstick, for example, in which adding one inch results in adding the same length no matter where the starting point is.

Another reason is that there is no one "right" scale in testing. The ACT and the SAT, for example, are reported on different scales. The way we would interpret a 5-point gain on a student's second attempt at one of these tests depends on whether the test is the ACT, reported on a scale of 1 to 36, or the SAT, reported on a scale of 400 to 1600.

3.1.1 Growth Compared to Similar Students

One way of providing additional information is to compare the growth of one student with the growth of other students – to provide normative information. Comparing the growth of one student with that of others allows us to make inferences about the relative magnitude of growth. When comparing the growth of students it is important to construct the comparison group with an eye toward the questions you wish to answer.

A trivial example would be to construct a comparison group of students with consecutive test scores in mathematics in third and fourth grade. This grouping would allow inferences about "typical" growth in mathematics from third to fourth grade. It wouldn't be helpful to include scores for students in other grades or for other content areas. We can take this even further, so that we compare the growth only of students with similar characteristics. These characteristics may include demographic (e.g., ethnicity), geographic (e.g., location in the state), or other test score characteristics (e.g., performance category).

3.1.2 Growth Percentiles

Betebenner (2008) has suggested that schools might find it helpful to consider a normative definition of growth that provides the average level of change from grade to grade for students with each starting point in their initial grade of testing. This approach results in what are essentially growth charts, very much like one finds in a doctors' office for height or weight. His idea regarding growth in school-related performance replicates the experience we have with our own children whose height is measured and whose growth can be compared normatively to other children. For example, in that context a parent might ask the child's physician: How tall is my child likely to be when he or she reaches age 18?

This same sort of question can be addressed to a school regarding a child's performance. How much increase in performance from year to year is normal for a child starting at a specific point on the performance scale? Relative growth can be studied normatively and each child can be compared to other children to see whether their growth is more or less rapid. For example, Betebenner provides a series of growth trajectories in *Norm- and Criterion-Referenced Student Growth* (2008) that relate to the students starting point.

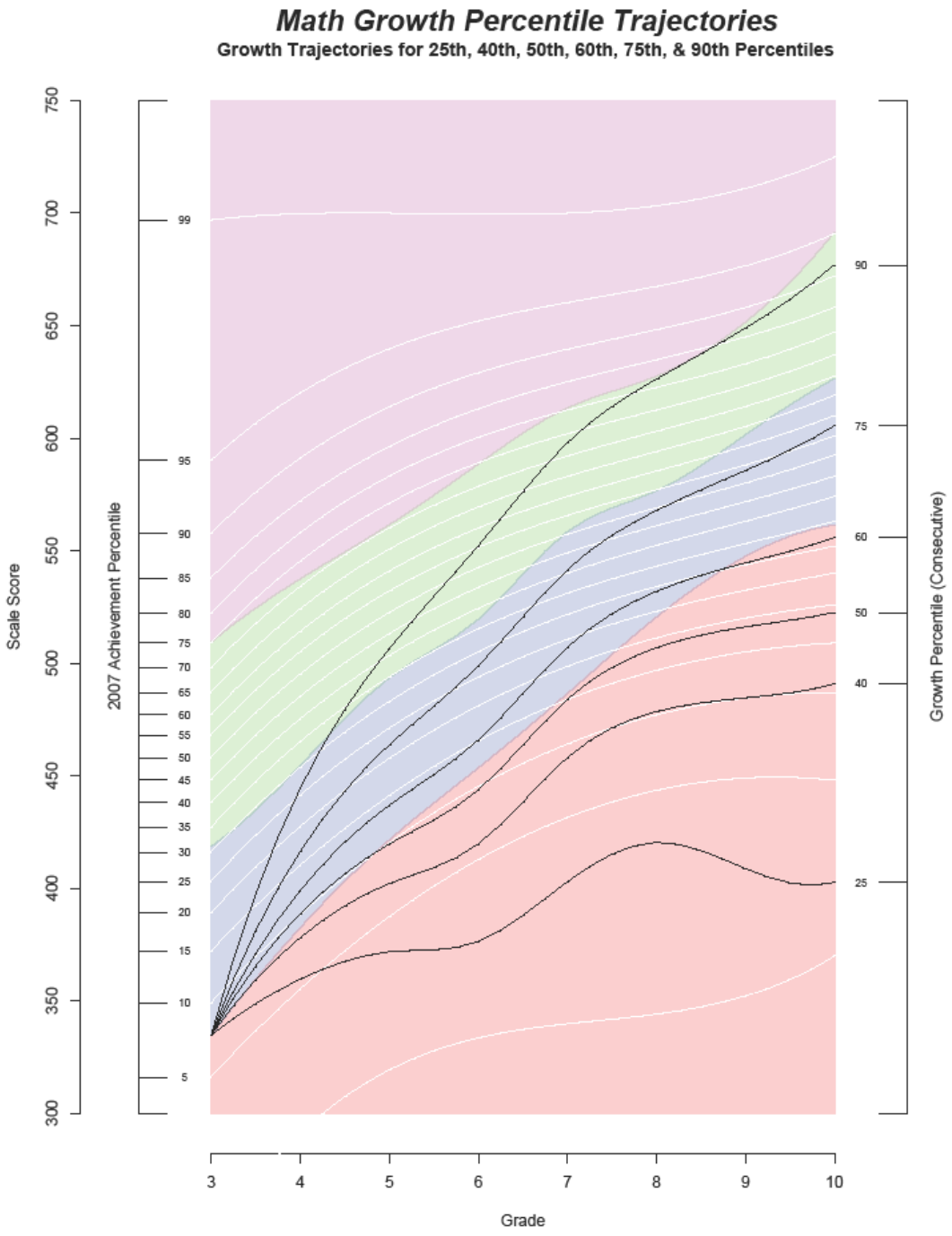

Figure 4. Sample mathematics growth percentiles (Betebenner, 2008). Used with permission.

Figure 4 depicts future math achievement for a third grader who is at the Basic cutpoint, conditional on 25th, 40th, 50th, 60th, 75th, and 90th percentile growth. On this chart the colored bands indicate performance levels, and the solid black lines indicate the growth

trajectory at each of the named percentiles. Grade levels are on the horizontal axis, and achievement is on the vertical axis. Note the vertical scale on the left. The chart indicates that for a third grader who is just Basic on this test (somewhere between the 5th and the 10th percentile), a growth rate better than 75% of all other third graders starting from the same point will be required to attain proficiency.

This kind of information can be very helpful in indicating what is reasonable to expect of a particular child in a school situation as well as what to expect from the school as a measure of success. For example, one can ask if a particular child with a certain score on a test in 3rd grade is likely to reach a certain point on the performance scale at a later time, say 8th grade. By modeling normative information on children with various starting points of performance, we can look at the distribution of their ending performance to see if reaching some particular level of performance is likely or rare. In terms of the height metaphor, we can describe the likelihood that a child will be tall or short given their starting height. Notice that in this conception of education, not every child is expected to learn as much as every other child.

Approaches such as Betebenner's may be a useful addition to models that look at what specific growth rate is needed to allow a child to reach proficiency in a certain number of years. These Normative models are essentially asking if various rates of growth are likely or not. This information can be a useful supplement to the other growth models already discussed above.

3.1.3 Technical Considerations for Normative Models

We described normative models as involving growth as a continuous variable, which is consistent with the typical model, but it does not have to be that way. For example, we could use the value-table model and think of that as a normative problem. What is the normal or typical change in performance category that we find in children? One can ask what percentage of children starting at the basic level in a grade move to the advanced level three grades hence? Again, this provides a base rate that might be helpful in understanding the variety of performance growth that occurs in schools. Some people think that expecting a school to move a child from below basic to proficient is not a realistic goal and such data can help a school system to understand that concern.

While normative information is important to determining realistic expectations, such an approach must be kept in perspective. Schools change, often for the better, and become more successful. When that happens the normative information should also be changed to reflect the new norms. Betebenner's approach does not have to be limiting. Benchmarking these normative data with other systems can be complex if the scale used in one system is not like that used in another. The effort to benchmark one school against another can still be an illuminating process in the effort to discover what might work better.

Finally, it is problematic to compare the growth rate of high performing students to that of other students given the limitations of test design (e.g., ceiling effects) or possible regression to the mean effects. It turns out that high performing students often have low or negative growth slopes while low performing students often have high growth slopes. Consequently, it is important when making normative statements about growth rates to only compare the growth rates of similar students.

In addition to providing information about relative growth, it may be very useful to provide standards-based information for a more complete picture of student, school, or district achievement. Standards-based and norm-based approaches look at achievement from different, perhaps complementary perspectives. With a standards-based approach, achievement is measured against a set of standards that is the same for all students, schools, or districts. These standards reflect educator judgment, policy considerations, and aspirations for achievement. On the other hand, with a norm-based approach achievement is measured relative to other students, schools, or districts (the reference group). Norms reflect a data-driven, descriptive approach in contrast to the prescriptive standards-based approach. Since students, for example, may show above average growth while still not meeting grade level standards, information from both approaches may be necessary for a more complete picture.

3.2 Value-Added Models

With the increased availability of longitudinal, student-level data (though not necessarily on a vertical scale) value-added modeling (VAM) has become quite prominent in educational evaluations. In contrast to the NCLB-specific growth models, VAM tends to be concerned with evaluating the performance of teachers, schools, districts, or programs irrespective of their performance vis-à-vis the proficient cut score. That is, under the framework of VAM, a school or teacher can be considered highly effective even when their students are not scoring at or above the proficient cut score. VAM can also be applied to assessing the effectiveness of interventions such as a new reading program or a special computer-assisted mathematics learning program.

In the state of Wisconsin, state statute prohibits teachers from being evaluated using test score data. For this reason, we generally orient the discussion around school evaluations without loss of generality. The value-added methods used for evaluating teachers are equivalent to those discussed below, but are of course at a different level of aggregation, focusing, for example, on the school, the department, or an intervention.

3.2.1 The History and Development of Value-Added Modeling

The concept of VAM as applied to educational data is credited to Dr. William Sanders, formerly at the University of Tennessee and now at the SAS Institute. The first published account of his methodology (Sanders, Saxton, Horn, 1997) revealed a groundbreaking technique for the analysis of educational data. While it remains controversial (Amrein-

Beardsley, 2008), the ideas spawned by Sanders have given educational analysts a method that portrays the quality of a school in light of the gains it makes with its students as opposed to the typical method that characterizes school quality via a single year of test scores.

The first implementation of his method occurred in the State of Tennessee under low stakes circumstances. In that state, *teacher effects*, a term we later define and consider, are estimated but are not used in any aspect of the teacher evaluation system.

In the years that followed, the idea of value-added modeling piqued the interest of other empirical researchers (Webster, 1997) and ultimately culminated in a special edition of the *Journal of Educational and Behavioral Statistics* (2004) that was fully devoted to the statistical methods used by VAM researchers.

This special edition can reasonably be referred to as the infancy of empirical VAM. The articles within that edition brought significant transparency to VAM and, subsequently, many papers on the topic have been published in that journal and others. Since that time, the literature on VAM has expanded to include new value-added models (Choi, 2007; Lissitz, 2005; Lissitz, 2006), computational methods for estimating the models (Lockwood, Doran, McCaffrey, 2003; Doran and Lockwood, 2005), and critiques of the models in general (Amrein-Beardsley, 2008; Kupermintz, 2003).

While Tennessee is the original VAM state, two more states have adopted VAM as a component of their school evaluation systems. The state of Ohio was required to adopt a value-added model per state statute (O.R.S. 3302.021). Pennsylvania has also moved towards statewide adoption of a school-based value-added model.

In both states, the VAM implementation is conducted through contract with Sanders at the SAS Corporation implementing the Educational Value-Added Assessment System (EVAAS); though in Pennsylvania it is referred to as the PVAAS. For clarification, however, the EVAAS value-added model is not the same statistical approach used by SAS for the NCLB approved growth model projections. There are two distinct statistical methods used by SAS. The most well-known is the value-added approach and the second is the NCLB projection model.

While VAM has not seen wide-spread adoption across states for teacher or school evaluation, it continues to be viewed as a popular and challenging research topic among academics. Recent national research conferences, such as the national conference on value-added modeling at the University of Wisconsin-Madison are illustrative of the wide array of VAM topics under continued scrutiny. Perhaps this scrutiny among academics is one reason that educational policymakers in state departments of education or school districts have delayed wide-scale implementation. Another reason may have to do with the complexity of the statistical analysis required to make inferences about the causes of student achievement growth. Professional development may be required for those who wish to interpret and use value-added results.

3.2.2 The Concept of Value Added

The purpose of this section is to provide greater transparency to value-added modeling (VAM) as a concept, and not as a method. In other words, our goal is to bring clarity to the reasons educators may be interested in VAM, what a VAM purports to accomplish, some of the controversial aspects of VAM, and ongoing related research topics.

There is some ambiguity in exactly what constitutes a VAM. This ambiguity has arisen, in part, because there are many different approaches to the problem with various statisticians presenting new methods (McCaffrey et al, 2004; Sanders, et al., 1997). Hence, there is no canonized, industry-accepted, definition.

For example, some analysts compute simple gain scores by subtracting a pretest score from a posttest score and refer to that gain as the value-added. Many of these simple models can be implemented in easy-to-use programs such as Excel. On the other end of the continuum, analysts implement complex statistical models that require extensive computing power. In these cases, highly specialized software is needed.

While there may be ambiguity in how a VAM is implemented, there is less uncertainty about what the purpose of a VAM is. In particular, a value-added model is a statistical method designed to determine how much *value* a school (or in some cases a teacher or a specific program) has *added* to a student's learning. Of course, this sounds simple enough, but implicitly includes an important assumption—some learning occurs within the boundaries of school and some learning occurs outside the boundaries of school.

VAM focuses on impact from within the school context and this assumption is critical. It clearly establishes the premise that schools should be held accountable for what they do, and not what occurs outside of their control. As such, the purpose of a VAM statistical model is to sort through all of the learning that has occurred (as measured by a student's improvement on a test) and extract the portion of that learning that occurred within the boundaries of school. The learning that is statistically identified as the portion that occurred within the school (or classroom or some special learning environment) is referred to as the *school effect*.

The VAM problem begins, as we said, under the assumption that only some of what students learn is attributable to what happens within a school. From that position, the intent of a VAM is to determine how much of an influence a school has had. Schools whose instructional programs have a significant influence on a student's learning are deemed effective whereas schools with little influence on what a student learns are deemed ineffective.

In the VAM literature, the portion of learning that occurs within a school or within a classroom is typically referred to as a *school* or *teacher* effect depending upon what the model is including in its equation. That is, the effect of having a certain teacher or the effect of attending a certain school is to exhibit learning gains as experienced by these students. In the scientific literature, this is referred to as a *causal effect* because we are

working under the assumption that some of the change in student achievement (whether it is good or whether it is bad) is caused by something that a school or district has done.

In much of science, such as in medical treatment studies, a causal effect is determined when units (e.g., rats, people) are randomly assigned to a treatment condition and others are assigned to a control condition where they may receive a placebo. For example, assume we have a new medicine to treat cholesterol and we want to determine if it can be used to reduce high levels of cholesterol in at-risk adults. In this case, our first task is to identify a group of at-risk adults.

We might begin with a preliminary assessment of the cholesterol levels in each of these individuals to serve as our baseline. Next, we would randomly divide these individuals into two groups: a treatment group and a control group. Those in the treatment group would be given the new medicine and those in the control group may receive a placebo.

After some period of time, we would reassess the cholesterol levels of all individuals. To determine the causal effect of the medicine, we would compare the cholesterol levels among those in the treatment group to those in the control group. If the cholesterol levels for those in the treatment group showed marked declines compared to those who did not take the new medicine, we might claim that this new medicine has an effect – to reduce cholesterol.

Now, turning our attention back to educational evaluations, we can relate the ideas of the preceding example back to evaluating schools. Our goal in using a VAM is to determine whether a school is having a positive effect on student learning or not. However, we are immediately under a constraint with school evaluations that is not found in much of science—randomization. Children are not randomly assigned to schools or even to classrooms. Parents choose their schools through a variety of means and principals may assign students to specific classrooms.

If students were randomly assigned to schools, there would be no need for a value-added model. Because the process of randomizing students to schools would "level the playing field" we could simply compare the end-of-year test scores across schools and determine which ones are highest.

However, we cannot compare end-of-year average test scores when there is no randomization because those scores are *confounded* with other characteristics of students. The term confounded is a term commonly used in research to mean that those scores are not the result of the school alone, but may be influenced by or impacted by other sources outside of the school. This confounding of scores with other non-school factors may unfairly give some schools an advantage over others and that advantage may have nothing to do with a school's instructional effectiveness.

The Cause of Student Learning

Proponents of VAM suggest that through the use of careful statistical modeling we can get an answer close to what an experimental design (i.e., a design with randomization of treatments to subjects) would give. The theoretical framework under which this occurs—referred to as the *potential outcomes perspective*—is described fully as it applies to VAM by Rubin, Stuart, and Zanutto (2004). We illustrate how this theory is used in VAM; however in doing so we must first construct a hypothetical situation.

Assume we have a group of 20 students and we assign them to attend school A in the fall of 2008. At the end of the school year, we test those students and examine their test scores. Now, how do we know if the instruction those students experienced was good or bad? Doing so requires a basis for comparison, or a *counterfactual*—a term that means how well these same students *would have done* if they had attended another school.

Now, for purposes of this example, assume it is possible to take these same 20 students, place them in a parallel universe, and simultaneously assign them to School B in the fall of 2008. Again at the end of the year we test the students and examine their scores. Under this hypothetical, we can compare the scores obtained for those students when they were at School A against those scores obtained when they were at School B. If the scores are higher at School A than School B, we would claim that School A is more effective than School B.

Of course, this scenario is impossible; we cannot simultaneously assign students to two different schools. But this scenario is similar to what a VAM is trying to accomplish. Through the use of a statistical method, we estimate two important quantities: a school effect and a comparison so that we can judge the school effect as relatively good or bad. Note that this approach is related to the desire to benchmark schools against each other to see which ones are more or less successful.

The VAM literature most commonly uses the average performance of students at the "typical" school for judging how well students would have performed had they been in another school setting. There are at least two methods for defining a typical school and these are both discussed below.

Reference Groups

In section 2, we showed how inferences regarding student growth trajectories are always made with respect to the performance categories. That is, we often asked if a student was likely to reach proficiency or not. As such, it is fair to characterize those models as standards-based growth models as the interest is always centered on whether students are growing toward a particular standard. In other words, these are criterion related models similar to what are called CRT approaches in testing.

VAM inferences, however, are different and are often conducted without regard to a performance category. The prior section discussed how the performance of students within a school is always compared to a counterfactual, or a reference group, in order to determine whether the school is high or low performing. In this regard, VAM inferences can be viewed as normative methods (the concept is similar to NRT assessment) for making inferences regarding schools as opposed to a standards-based model.

In the GTS approach, the performance category selected can be from any future grade. For instance, the examples given earlier showed how inferences regarding student growth trajectories were made with respect to the grade 7 standard. There is nothing sacrosanct about this choice—any other grade for which appropriate data exists could have been chosen. In the same spirit, one must also consider the reference group that will be used when making inferences regarding schools from a VAM. In general, there are two approaches to selecting a reference group that are used and we discuss those two here.

District as the Reference

One approach to implementing VAMs is to analyze the data on a district-by-district basis. In other words, all data from District A are analyzed independently from the analyses performed for District B and so on. When this occurs, all schools included in the analysis are compared to the average unit in that district. Because the analysis is performed in this manner, it is not possible to compare the value-added school effect in Madison to the value-added unit effect in Milwaukee, for example. Schools in Madison can only be compared to other schools in Madison and schools in Milwaukee can only be compared to other schools in Milwaukee.

There are a few reasons this approach may be chosen. First, from a technical perspective the computational burden is largely reduced because the data set for a district is smaller than analyzing all the units in the state at one time. It is usually much easier to analyze smaller data sets than it is to analyze larger data sets.

Second, when districts implement a VAM on their own, they are likely not to have access to student-level data from districts other than their own. That would preclude a state wide analysis.

Third, districts may prefer a customized VAM that addresses their particular needs and capacities. Districts may have data that are not available at the state level, such as benchmark assessment scores or program- or classroom-level information. A common question over which districts may differ is whether to include student background characteristics such as socioeconomic status or race/ethnicity as covariates in the model (see the discussion of validity in section 3.2.3 for more on this).

State as the Reference

Another approach that may be chosen is to perform the VAM analysis using the data for all schools in the state concurrently. When this is conducted the performance of a school is compared to the average school in the state. As such, schools in Madison can be directly compared to other schools in Milwaukee.

With respect to statewide accountability systems, it would seem that choosing the state as the reference would make most sense if the data can be analyzed as such. An analysis of this form allows for a direct comparison of all schools within a state and this is typically the goal of a statewide accountability system. Again, being able to perform this analysis depends upon having the relevant data from the whole state.

3.2.3 Technical Considerations for Value-Added Models

The current body of research within the field of VAM is rather extensive. A review of the papers at http://www.wcer.wisc.edu/news/events/natConf_papers.php gives an indication of the topics covered at the national VAM conference in Wisconsin (Wisconsin Center for Education Research, 2008) and that are representative of the ongoing research in this area. The next few sections highlight the salient technical issues in VAM, with the intended goal for the reader to be aware of the issues presented and have an understanding of how these issues may affect the choice of a value-added model.

The Effect of Prior Schooling

We begin by contrasting two of the more prominent value-added models proposed in the educational literature: EVAAS and the RAND model. The EVAAS model was previously described and is sometimes more commonly known as the "Sanders" model. The RAND model was first proposed in the special edition of the *Journal of Educational and Behavioral Statistics* (2004) and offered an important alternative that we consider here.

The EVAAS model is also sometimes termed the *layered* model. The term layered in this context is used to mean the effect of the prior school or teacher remains with the student entirely even as they progress through the school system. In many respects, this is akin to assuming that students retain all of the information they learned in the prior school years.

Researchers at the RAND Corporation considered layering not to be completely plausible. Their view is that the effect of the prior school (or teacher) may not make the same contribution to the current school year estimate as is assumed by the EVAAS system. Consequently, they proposed an alternative value-added modeling approach, the *persistence* model that allows for the contribution of the prior school to be estimated from the observed data and not fixed as a constant as is assumed in the EVAAS system.

From a pragmatic perspective, this is an important issue to consider when choosing a VAM, for two reasons. First, the estimates of the school effects will not necessarily be the same between the two models. Second, the RAND model is slightly more complex than the EVAAS system and there is currently no publicly available software that can estimate school effects from this model.

Random Effects versus Fixed Effects

As VAM matures, at least two methodological pathways have emerged: random-effects and fixed effects. This section describes in lay terms some of the important characteristics of these two methodologies.

It is often the case that analysts consider many *factors* that affect student learning. Some of those factors, such as gender, are dichotomous in that the factor includes only two *levels*: a student is either male or female. Other factors, such as ethnicity or school have multiple levels. For instance, the levels of race may include American Indian/Alaska Native, Asian/Pacific Islander, Black, Hispanic, and White. In this case we say that race has five levels. Schools are also a factor with as many levels as there are schools in the study.

When a researcher implements a value-added statistical model, they must choose whether to implement a model with *fixed effects* or a model with *random effects*. This decision typically hinges on whether the factors included in the statistical model will expand during the course of the study. Another way of stating this is whether the factors included in the current study can be viewed as a sample from the population of factors under study. If the factors will not expand (or are not viewed as a sample), it is said that they are fixed and a fixed effects model may be chosen. If the levels of the factor will expand, then it is said that the factor only represents a sample from the population of that factor.

For example, gender is a factor that will never expand during the course of a study. Students can only assume one of two levels and those levels will never change. In this case, gender should be modeled as a fixed effect. Schools or students, on the other hand, will change during the course of the study. New schools and new students will emerge each year in a value-added design and should be modeled as random effects.

Now, in some cases, a statistician creates a statistical model that has a combination of fixed effects and some random effects. When a statistical model includes both, it is commonly referred to as a *mixed-effects* model in that it combines both the fixed effects and the random effects into the same statistical model. In contrast, when all factors are treated as fixed, the statistical model is referred to as a *fixed effects* model.

It is at this point where a rather substantial difference appears in the VAM literature (Meyer and Sass, 1993). Some VAM analysts view students and schools as random effects and implement a mixed-effects model whereas others view students and schools as fixed and implement a fixed effects model.

One might reasonably wonder whether such a difference would have any consequence on the results. The answer is "yes"; there is a substantial difference between the results of a statistical model that views schools and students as fixed and one that views students and schools as random.

While the rationale for these differences is best described via a mathematical conversation (Doran, 2008), it boils down to a rather simple concept. When a value-added result is derived for a school from a fixed effects model, that VAM estimate is derived only on the basis of those students that are "movers" into the school whereas in a random or mixed effects model, the VAM estimate is derived on the basis of all students attending the school. In other words, a fixed effects model uses test scores only for students that are new to the school.

Whether or not this issue will add some bias to the school effect depends on the characteristics of those new students. If the students that are new can be reasonably viewed as a random sample of the school's population, then it is likely that the VAM is not very biased for that school. However, this tends not to be the case. New students tend to be more transitory and this is often associated with some characteristics of the student. For example, students with lower economic status tend to move into and out of schools more often than students with a higher economic status.

Further compounding that issue, if a school has no new students (or very few) then the VAM estimate cannot be determined. For example, assume we have a school with 500 students in it. Next, assume none of those students are new to the school. It would then be impossible to estimate a VAM for that school if a fixed effects model were used.

Researchers also use the idea of *nested factors* in hierarchical linear models (HLM). Factor A is said to be nested in Factor B if every level of Factor A is observed within only one level of Factor B. For example, assume students constitute Factor A and assume schools constitute Factor B. Students are said to be nested in schools if and only if every student test score comes from one and only one school.

With cross-sectional data (i.e., when students have only a single test score) this nesting structure is easy to satisfy. However, VAM methods are longitudinal and students have more than one test score used in a VAM analysis. For many reasons, students move between schools over time. As a result, the first test score for a student may be obtained from a different school than the second test score. When student test scores are obtained from more than one school, the nesting structure of the data is no longer in place and HLMs no longer suffice.

Psychometrics and VAM

An important, yet smaller, body of research within the world of VAM has examined the issues associated with item response theory (IRT) and whether vertical scales are

necessary for modeling growth. Ballou (2008) has recently argued that IRT scaling methods do not produce truly interval scales; a property that many of the statistical models used for VAM require, although not all of them. Instead, he argues that at best IRT results in an ordinal scale and yet many current VAM methodologies do not treat the data as ordinal.

Martineau (2006) argued that IRT scales are not unidimensional but are instead multidimensional and ignoring this dimensionality can have serious effects on the estimates derived from a VAM. Doran and Cohen (2005) have argued that the linking error that arises as a function of linking different test forms adds to the variance in the test scores, but that this error is commonly ignored.

Briggs, Weeks, and Wiley (2008) examined various methods of vertical scaling and their impact on value-added results. Using student-level data in Colorado, the researchers created eight types of vertical scales which they then used to compare school-level effects. In this study, they found a strong and positive correlation between school effect estimates between the eight vertical scales. Results indicated large numbers of schools switched categories when using different vertical scales, but the percent of schools classified as high or low performing remained the same regardless of vertical scale used to categorize the schools.

Validity of VAM

It is noteworthy that to date there is no published experimental validity research on the topic that can suggest which VAM is "right." That is, there are multiple approaches to VAM, and those approaches are likely to give different estimates of school quality. However, which of those results provides the most accurate answer is still unknown.

Rubin, Stuart and Zanutto (2004) argued that this is a difficult, if not impossible question to answer. Rather, they suggest that the more appropriate question is to investigate whether schools make use of the results of a value-added model to cause instructional changes to occur within the schools. Responding to this argument, McCaffrey and Hamilton (2007) conducted a comprehensive study in the state of Pennsylvania and found that the use of the value-added results by teachers was very limited and that there was no difference in achievement between schools using the PVAAS system compared to those schools not using PVAAS.

The implications of this study are not as grim as they may appear. Indeed, the authors of the study even note that because teachers in this state are not held accountable based on the results of the value-added model, it may be information that is viewed as less useful than other test score results, such as the AYP information generated for NCLB.

If teachers and principals are not specifically trained to interpret and use assessment data, it is not surprising that statistical information is not used to make changes in the

interventions for which they are responsible. It is a simple fact that people cannot effectively use information that they do not understand.

Kane and Staiger (2002) conducted an experimental study to evaluate value-added estimates. Using data from previous years, they predicted teacher performance for a large sample of teachers. Teachers were then randomly assigned to classrooms of students and data were evaluated to see whether value-added gains were accurately predicted. The researchers came up with two major conclusions: First, they found that they were able to use non-experimental specifications (such as background characteristics of students) to predict the outcome; and second, they found differences in the impact teachers had on math performance versus English language arts performance. They suggest conducting similar studies at other districts to provide better support for their conclusions.

Issues of data quality, which include decisions about what variables should be measured, are. The paper by Lissitz, et al. (2006b) discusses several very important issues regarding the selection of variables to include in a VAM model. If a variable is not included in the model, it will not be considered in the analysis. For example, Adcock (1995) found that teachers with advanced degrees seem to be somewhat more effective than teachers with bachelor degrees. A school system has a choice whether to put teacher education level into the model that they want to develop to try to understand what seems to cause greater growth. They could include race and gender too, but if they do so they must recognize their model will differ from what is currently allowed for statewide accountability under NCLB .

The pressure from NCLB is one reason not to include such variables, but Lissitz, et al. (2006b) also discuss another reason that is probably more important. This has to do with what the school wants to do with its findings. If they are simply trying to identify characteristics that are associated with highly successful students then they might consider many variables that are predictive of performance. If, on the other hand, they are looking for variables that might provide ideas for interventions by the schools, then such demographic variables such as race and gender that can't be changed would not satisfy that desire. In that case, as Lissitz, et al. (2006B) suggest, the model should only include variables that have the potential to be manipulated and changed to the benefit of the student's performance. As these models get more and more sophisticated in the future, such questions become more and more important for the schools to agree upon.

Software Requirements

Throughout this paper we have noted that specialized software is needed to estimate the more complex value-added models discussed. This problem was first discussed by Sanders, Saxton, and Horn (1997) in the context of the TVAAS model and was discussed by McCaffrey, et al. (2004) as perhaps the greatest barrier to VAM implementation. Lockwood, McCaffrey, Mariono, and Setjodi (2007) further discuss this issue and provide a possible solution, although that solution still requires an enormous amount of computing power and possibly many days of time to complete the computations even

with high speed computers. Doran, Bates, Bliese and Dowling (2007) document a computing method that significantly reduces the computational time and may prove to be a promising method for others to replicate.

The fact that specialized software is necessary to implement VAMs is not trivial nor is it necessarily a call to reduce the complexity of VAM models so that implementation is easier. Many of the VAM methods proposed in the literature are computationally complex because they are aimed at reducing the amount of statistical "noise" in the data in order to squeeze out the most reasonable estimate of causes of student performance. In other words we are trying to see if, after taking into consideration differences between students in their overall achievement level and their growth trajectory and various factors that apply across the population (gender, race/ethnicity, socioeconomic status) we can discern effects for schools or for teachers beyond what we would expect from random chance.

Vertical Scale

A vertical scale is one in which performance can be placed on the same scale across several grades. In other words, if we have a vertical scale we will be able to compare the performance level of a student in third grade to the performance level of that same student at 4^{th} or 5^{th} or grades beyond that. It is a scale that behaves a little like height or weight and has a developmental quality about it, since for most children they obtain higher scores as they progress through school. We can compare the height of a 3^{rd} grader to that of a 6^{th} grader because height is measured on a very straightforward form of a vertical scale. In assessment we can also create such a scale.

There are other very simple models that states can use to look at growth that assume stability of the model each year, but do not involve careful comparison from year to year. Schafer (2006) presents a very simple model for doing so. As with everything else in life, schools will need eventually to choose between simplicity and sophistication, recognizing that sophistication usually requires more money and effort. The point is that a state can decide to do its testing in a way that facilitates the creation of a vertical scale, or the state can choose not to do so. Not doing so limits some of the other decisions that a state can make, but in fact there are other ways to answer many questions of interest that do not require a fully developed vertical scale.

3.3 Summary of Models for Research and Program Evaluation

Value-Added models seem to present some promise in school evaluations. They are, of course, not without controversy and the field is constantly changing with new developments. However, it has been noted that one "cannot allow the perfect to be the enemy of the good." Status based models, as simple as they are to compute and as easy to explain, are controversial as well given that the distribution of student abilities across schools is unequal making status models very difficult to interpret correctly. Comparing

end-of-year test scores with no attempt to control for those pre-existing differences is an unfair and misleading comparison. VAM, on the other hand, is difficult to explain, but does explicitly attempt to control for those pre-existing differences among students across schools.

There is no gold standard by which one can choose a statistical method. That is, we cannot know for certain whether the layered model, the persistence model, a model with fixed effects, random effects, or a model with covariates is the right model. However, each model comes with certain assumptions and Wisconsin school systems considering implementation of these models should consult with statisticians who can likely offer advice on which model makes most sense given the structure of their data and the assumptions they believe are most realistic as well as the purposes they have for adoption of such models.

4. Recommendations for Wisconsin

We begin this section with questions to consider when deciding how to model growth, including some considerations specific to Wisconsin. We conclude with detailed descriptions of models for further investigation in Wisconsin.

4.1 Questions to Consider

This white paper has talked about a number of models that address issues of growth and the determination of value added. In this section, we want to talk about some of the issues that Wisconsin might consider as they think about which of these options, if any, they will choose for implementation. As the reader will remember there are several models that are currently used to measure growth that have been accepted for NCLB accountability. If we ignore the subtleties of each model, we can identify two basic models that have been approved and these are 1) growth towards the standards and 2) value tables. Different, but related is the 3) normative approach which we indicated could be considered as a supplement to student specific growth models. Finally there is the family of models we called 4) value added which, although the federal government does not currently accept them for accountability purposes, have been implemented by some states.

The process for choosing a growth model cannot be laid out in a simple decision tree, but there are certainly some considerations that must be resolved and doing so will guide the selection or creation of a model or models. Considering sample analyses is also an important part of the decision making process and will be done and reported in the final, technical paper for this project, but here we want to suggest some general issues that might influence Wisconsin. These are identified to motivate Wisconsin to engage with their various stakeholders as they consider how to move forward to develop a balanced scorecard approach or what we call education science.

Table 6 and the following sections discuss the questions educators can ask themselves as they try to decide in which direction they wish to go. Note that the process may actually be iterative, as states acquire more expertise in answering certain essential questions, rather than linear, as depicted here. We conclude with a summary of the questions for consideration in Table 7.

Table 6. Decision process for choosing a growth model

You must consider this		Before you can do this
Inference: *What do you want to know?* Capacity: *What data and resources are available?* Policy: *What values will guide your decisions?*	⇨	Select or create a growth model.

4.1.1 Inference

What do you want to know?

Which of the typical parent, teacher, and administrator questions from section 1 are of greatest interest? Do you want to know if students are on target to be proficient? If a certain intervention is effective? If one school is as effective as another school? To a great extent the questions of interest will guide model selection. For example, if you want to know whether a child has made "a year's worth of growth," you will need normative information. On the other hand, if you want to know if a child is growing enough to meet state standards, you might choose a growth-towards-the-standards model, or combine standards-based and norm-based information in your model.

NCLB Accountability

Does Wisconsin want to adopt a model that is approvable under current NCLB requirements? For example, if you want to use demographic data to characterize the rate of growth, doing so is not currently acceptable for NCLB accountability. There is a certain risk to policy makers in adopting analyses that are prohibited by the USED. If use for NCLB accountability under current requirements is important you will want to select either a value table or a growth to standards approach.

If Wisconsin adopts an NCLB-approvable growth model, a further question is at what point in adequate yearly progress determinations the growth model should be used. When growth models are implemented for purposes of NCLB accountability, states must decide

where in the decision making process the growth model results are to be used. For example, a state may implement their adequate yearly progress (AYP) calculations as follows:

1. Compute status results.
2. If the school fails to meet the status AMO, apply a confidence interval.
3. If the school fails to meet the status AMO with a confidence interval, apply uniform averaging.
4. If the school fails to meet status AMO with uniform averaging, implement safe harbor.
5. If the school fails to meet safe harbor, implement a growth model calculation.

However, a state could just as well first implement the status approach, then the growth approach, and then all the other methods used for the AYP calculations. As it currently stands, there are no statutory or regulatory provisions that would preclude a state from moving growth to a more prominent position in the decision making process for determining AYP.

There have been some claims from states that growth models didn't matter in their AYP process (i.e., few, if any, schools were reclassified as having made AYP because of the growth model). When a state has multiple methods for determining AYP (e.g., status, safe harbor, confidence intervals) and growth is but one method, this result is not entirely surprising. This is largely a function of the fact that all of the other methods in the AYP decision making process, in combination, can have identified most of the eligible schools. Hence, when a growth model is added, there may be little additional change. This makes the claim that growth doesn't matter for AYP (i.e., doesn't matter in the sense that it doesn't reclassify schools as making AYP), something of an unfair argument.

Another consideration, as we indicated at the beginning of this paper, is that growth modeling can answer questions that can't be answered with status models and the advantages to answering such questions seem to be potentially great. Moving beyond AYP decisions holds great promise for schools and states that want to adopt a balanced scorecard and an education science agenda.

Teacher, Program, School, and District Effects

If Wisconsin adopts a VAM model, you will need to decide which plausible causes you want to study, if any. Do you want to be able to evaluate school or teacher effectiveness? What about special programs such as a reading program or a mathematics instruction module? Perhaps you are interested in the effect of a student's choice of electives? Does the state data system include the requisite variables? If so, a value-added model is going to be your choice.

4.1.2 Capacity

What data and resources are available?

Wisconsin has developed a vertical scale already, but is it sufficient to be used for growth and/or value added purposes? If so, interpreting growth with a good vertical scale is in many ways easier to explain because of the analogy to growth in height. Even growth models can be implemented without a vertical scale as long as there is reasonable stability in the interpretation of the scale at each grade level. In other words, the scale at each grade must be stable from year to year, neither getting harder nor easier as the testing is repeated across years.

Wisconsin has unique student IDs that can be used to create a longitudinal data structure. The state has four years of data in grades 3-8 in mathematics and reading. However, students are tested only once in high school, in grade 10. Some students take an alternate assessment, which is reported on a different scale than the WKCE. The state longitudinal data system includes many individual student variables, but no information tying students to classes or teachers.

The availability of computer programs, people to run them and especially people to interpret and explain them, can become an issue particularly if Wisconsin decided to adopt a very complex model. If this worries Wisconsin, you may want to avoid the more ambitious VAM models.

Standards and Assessment

When modeling growth the technical quality of academic standards and assessments matters in new ways. For achievement growth to be meaningful, the academic standards must represent a clear developmental continuum. For test scores to be useful without additional statistical manipulation (such as normalization), test forms must be the same difficulty and measure similar content and skills each year. This is referred to as *horizontal equating*. When two different test forms are horizontally equated we can infer, for example, that a child scoring 450 on Form A in year 1 has about the same level of knowledge and skill as a child scoring 450 on Form B in year 2.

If this is not the case, it is impossible to know whether score scale differences in student growth from one year to the next are a result of real learning or simply due to a test form being relatively easier or harder in one year than in another. In other words, we couldn't be sure that a child whose scale scores changed from 450 in third grade on Form A to 480 in fourth grade on Form C actually showed the same amount of growth as a child whose scale scores changed by the same amount from third grade on Form B to fourth grade on Form D. We couldn't know how much of the 30-point gain represents learning and how much represents "form effect" – differences in the difficulties of the forms. If Forms A and B are the same difficulty, but Form D is actually much more difficult than Form C, then the second student's 30-point gain represents more learning.

When a vertical scale is used, it is important to know that the scale was carefully constructed. It is also important to understand the strengths and limitations of the scale. For example, static tests designed to assess proficiency for NCLB purposes may measure student knowledge and skills very accurately around the Proficient cut score (the minimum score required to be classified as Proficient), and much less accurately at the extreme ends of the scale. This makes such a scale less reliable for measuring the growth of either very high- or very low-performing students. One solution – apart from having a much lengthier test – is to use an adaptive test format.

Professional Development

Many teachers and administrators will require additional training to effectively use the results of value-added models in data driven decision-making. Does Wisconsin have the desire and the resources to get involved in this kind of professional development?

How will Wisconsin handle training educators? Will it be through direct training or development of supporting documentation? Who will provide the back-up expertise? Do the schools have such personnel? Does the state or other larger districts such as Milwaukee have such expertise that they can call upon? If this worries Wisconsin unduly you may decide to either go slowly or not go in this direction at all. Perhaps a federally approved growth model would be the most conservative initial approach.

4.1.3 Policy

Normative Information

Does Wisconsin want their results to be compared to a base rate? If so, do they want a normative value to be defined for schools with very different students or should the same comparison be demanded for every school or system – what will be the reference group? If a base rate is important to Wisconsin educators, you might want to develop normative growth tables so that everyone can see what "typical" growth is. Several of these could be developed for use by schools that differ significantly in their characteristics.

A normative growth model might be a useful supplement to adopting a growth to the standards model. See sections 3.1 and 4.2.2 for additional discussion of using normative information.

Inclusion of Student Background and Other Information

Various demographic variables might be considered such as SES, percent minority students, number of years of experience of average teacher, or percent of teachers with MA or higher degree. There is some feeling that including these variables may tend to

create different expectations for different students. On the other hand, not including some of these variables may reduce the ability of the model to determine program effectiveness. See section 3.2.3 for more information.

Focus on the School, the Student, or Both

The value-table approach focuses upon the school level for its results. Individual students are not identified, nor are individual teachers usually identified, although it will become clear eventually whose students are contributing positively to the outcome level summarized in the value-table calculation. In the approach using a value table the key performance question revolves around the school for which the performance category data are aggregated. Debating, for example, about how much a school values a move from proficient to advanced, or from basic to proficient can unify a school around its goals for its students. The success of the school becomes the focus versus the success of individual students or classrooms.

If Wisconsin wants to have each child assessed for his or her likelihood of reaching a target, then the value-table approach is not going to be acceptable. Student focus suggests a growth to standards model and not a value-table approach. If selecting a growth to standards model, a further decision that is required is whether to include all students, even those who are already at or above proficient.

Summary of Questions to Consider

Table 7 provides a summary of these questions and issues to consider. In any case, we recommend that Wisconsin educators collaborate with researchers to explore modeling possibilities before operational implementation occurs for accountability, evaluation or any other purpose.

Table 7. Summary of Issues to Consider When Choosing or Creating a Growth Model

Inference
What do you want to know? *NCLB Accountability* *Teacher, Program, School, and District Effects*
Capacity
What data and resources are available? *Standards and Assessment* *Professional Development*
Policy
Normative Information *Inclusion of Student Background and Other Information* *Focus on the School, the Student, or Both*

4.2 Recommended Models for the Wisconsin Project

There are two overriding goals for this project. The first is to create growth models that fit the Wisconsin data. The second is to create models that allow evaluation and comparison of the major approaches to looking at growth. In conversations with WDPI we considered several NCLB growth models as well as ideas for exploring technical aspects of value added models.

The data available to us include results from three years of WKCE testing (fall 2006, fall 2007, and fall 2008) in two content areas, reading and mathematics. For each student we have demographic information (age, gender, race/ethnicity, economically disadvantaged status, disability status, and English proficiency code) and information on the school and district attended at the time of the examinations. Test scores are available as scale scores, raw scores, and performance levels.

We recommend that Wisconsin consider the following four models. First, the state has expressed an interest in a simple growth model. The state has also requested a normative model. The term normative is used to mean that a student's growth is compared to another growth rate, rather than to a criterion such as the growth required to achieve proficiency at some future point in time.

The third model, the probability of proficiency model, could be used as part of adequate yearly progress calculations under NCLB. This model was developed after reviewing the Wisconsin data. Our explorations with the data suggest that the model works well and can be used for both NCLB accountability and value-added research or program evaluation-related decisions. The fourth model is a value-added model that would not be allowed for NCLB accountability but has been deeply considered in the statistical literature and has been implemented in other states. However, it has never been implemented to a full state-wide data set given certain computational limitations. We can overcome these limitations and can fit the models for the first time using all state data.

4.2.1 Simple Growth

The simplest means of modeling growth – subtracting last year's scale score from this year's scale score – is one that can be done easily on a spreadsheet without sophisticated statistical modeling. It is easy to understand and report. However, as noted above (section 3.1) it suffers from important limitations. In view of WDPI's expressed interest, we recommend including a simple growth model as a basis for comparison with more sophisticated models of growth.

4.2.2 Normative Growth

With a normative model the goal is to compare growth for one student against growth for other, similar students. We propose to implement a "similar student index" that allows us

to create a comparison group of some number of students who most resemble one another in some specifiable ways, such as prior test scores or geographic location. We can easily implement an algorithm to accomplish the following:

- Identify a set of *N* students in the data set that most resemble student *i* given a set of characteristics.
- Compare the growth rate of student *i* to that of the *N* other students in the data most like student *i*.
- From the comparison, generate a set of growth norms, such as percentiles.

Given a set of similar students, we can make statements such as, "The growth rate for this student is at the 75th percentile when compared to 10,000 similar students." It may also be of interest to compare a student's growth rate to the mean growth rate for all students in the comparison group, allowing us to make statements such as, "The growth rate for this student is above the average growth rate when compared to 10,000 similar students." Both of these statements are normative comparisons.

4.2.3 Probability of Proficiency

WKCE scores are reported using a vertical scale, which allows the use of a growth-towards-the-standards model for calculating adequate yearly progress (see section 2.3.1 for more information on these models). Such models tend to be concerned with the following questions regarding students and schools:

1. Given the observed scores for all students in a school this year, what proportion of these students are likely to be proficient next year?
2. Given the observed performance of a particular student in a school this year, what is the probability that the student will be proficient next year?
3. How does the performance of students in one school compare to the performance of students in other schools, after controlling for differences in student abilities?

NCLB accountability is primarily related to the first question; classroom educators and parents tend to be concerned with the second; and school administrators and policymakers are often concerned with the third (comparisons of "school effects" or value added). A useful model would attempt to answer all three questions at the same time in a reasonably simple, yet reliable way.

We propose a "probability of proficiency" model motivated by three assumptions:

1. Many of the growth-towards-the-standards models proposed for the USED growth model pilot program project future performance as a fixed score, known with certainty. However, future success cannot be accurately projected, only estimated with some probability.
2. The probability of proficiency in the second year is related to a student's score in the first year. Students with high scores in the first year have a high probability of

proficiency in the second year, whereas students with low scores in the first year have a lower probability of proficiency in the second.

3. Schools have a differential impact on student performance. That is, some schools are more effective with their students than others in terms of growth. As a consequence, two students with the same scale score, but attending different schools, may have different probabilities of future success as a result of differences in the instructional programs.

Given these assumptions, the model we propose generates a probability that a student will be proficient next year based on their school and scale score this year. We calculate this probability for all students, whether or not they are currently proficient. We do this because it is possible that a student who is proficient this year would not be proficient next year. Our aim is to include all students, not just those scoring below proficient.

We focus on predicting proficiency, rather than predicting scores, to improve the validity and reliability of the predictions. The validity of any statistical model for a response in year 2 as a function of a score in year 1 will be limited by the uncertainty of the score in year 1. The score in year 1 is an imperfect indicator of a student's ability that year. Issues of repeatability and vertical or horizontal equating must always be considered when using such scores.

However, if we focus on modeling the probability of proficiency in year 2, as measured by the observed proportion proficient, we can diminish some of the undesirable effects of the "pre" score being an imperfect measure of ability. The reason for this is that the points that are most influential on the curve predicting probability of proficiency are the points where the test most accurately measures ability. With a fixed form test such as the WKCE, there is typically more measurement error at the extremes of the ability scale than in the middle. This is because the test is designed to be most accurate near the proficiency cut score than at any other point along the scale.

For calculating adequate yearly progress we can assign a probability of proficiency to each student in a particular grade in a school based on past performance of students and the scores of the current students in the school. The sum of these probabilities is the expected number of proficient students in that grade on the current year's test (see table 8 for sample calculations). After the test has been administered and scores calculated we can compare the observed number proficient to the expected number.

The current NCLB regulations do not allow the use of demographic variables or specifics of past performance in the particular school to be used in formulating the expected number of proficient students. We do have a choice of reference group in that we can formulate the model from the state-wide results or from district-wide results.

Table 8: Sample Grade 3 to 4 Projections

Number of Students	Score in Grade 3	Number Proficient in Grade 3	Probability of Proficiency in Grade 4	Number Projected to be Proficient in Grade 4
10	400	0	.18	2.4
20	420	0	.52	10.4
15	430	15	.70	10.5
6	440	6	.84	5.04
8	450	8	.92	7.36
Total		29 (49%)		36 (61%)

We can also make student-specific predictions to provide parents, teachers, or administrators with the most accurate estimate of the probability that a given student will be proficient on the next test. In this case we would incorporate demographic variables and information on the school and district. The model could be fit to state-wide results or to results from a particular district. The advantage of using state-wide results is greater precision. A district-specific model may be more appropriate for some of the larger districts, although it would be hopelessly imprecise for small districts.

In order to assess which schools are doing significantly better or worse than a typical school in raising students to proficiency, we must include information on the school in the model. The reference group of schools could be all those in a district or all those in the state. Demographic variables can be included or excluded, depending on whether we wish to control for the characteristics of the student body in assessing the effectiveness of individual schools.

The benefits of this model are many. First, every student with an observed test score can be included when forming the projections. Second, the data only offer the chances that a student will be proficient in the subsequent school year rather than assuming that a projected score is known with certainty. Therefore, this model is conservative – it makes no claims beyond what the data can support. However, the probabilities are instructionally useful at the student level because they may provide a call to action for some form of instructional remediation in an attempt to improve a student's likelihood of success in the following school year. Finally, the model can be used to determine which schools are producing gains in student achievement that are larger or smaller than other schools in the state.

4.2.4 Mixed-Effects Value-Added Model

Whereas the probability of proficiency model is clearly focused on determining whether students are "on track" to reach the proficient cut point, a value-added model is more concerned with determining the "effect" of a school on student achievement. The properties of a mixed-effects model are well known – it is perhaps the most widely-used value-added model, and it has appeared widely in the literature.

A mixed-effects model uses the entire set of achievement scores for each student and makes specific assumptions about the extent to which the impact of prior schools remains with individual students. In general, as discussed in section 3.2.3, there are two competing assumptions regarding the impact of a prior school. The first, often referred to as cumulative school effects (or "layering") and stemming from the Sanders model, assumes the impact of a prior school experience persists at a constant level for each student. The second assumption, which stems from the work of RAND, is that the impact of a school diminishes with time in a way that can be estimated from the observed data.

Value-added models typically include covariates such as student, school, district, year, grade level, demographic information, and so on. Some covariates included in the model are best modeled as "fixed effects" – such as gender – for which the set of possible levels is fixed (in this case, male and female). Other covariates are best modeled as "random effects" for which the set of possible levels will continue to expand as the study progresses. For example, the sets of students, schools, and districts that we observe will change over time. Mixed-effects models use both fixed and random effects.

In the past it was important to distinguish models in which the factors for the random effects were nested from those that did not have this property. A factor like "student" is said to be nested within "school" if each student is observed in one and only one school. In such a case we can consider the data as being modeled first at the level of the school, then at the level of student within school, and then at the level of observation within student. This led to the development of multilevel, or hierarchical linear models. There are certain computational simplifications that apply to models with nested factors for the random effects. Current methods of fitting mixed-effects models do not require these simplifications. We do not require nested factors for the random effects.

5. Conclusion

The testing requirements of No Child Left Behind have generated an opportunity to implement methods that analyze data longitudinally. While this opens up new possibilities for state accountability, it also requires a significant amount of consideration of the multiple issues involved in developing and implementing growth models.

Many of the proposed growth models have not been fully vetted by researchers for quality nor have they been in place long enough to demonstrate that they are well-suited for state accountability and/or evaluation. Even so, modeling growth in student achievement is likely to offer significant benefits in understanding school and teacher quality. Status-based approaches utilizing only a single test score will always have the flaw that those scores are confounded with other non-school factors. Growth models have the potential for disentangling school effects from other non-school factors, at least to some extent, in order to improve the quality of judgments that can be made regarding school quality.

It is also unlikely that growth models will assume the same role as item response theory (IRT) has in educational measurement. For instance, test publishers and states routinely make use of the same IRT methods: Rasch, 2- or 3-parameter logistic models, or the generalized partial credit model for polytomously scored items. However, it is unlikely, and probably unwise, to assume a set of specific growth models or approaches to VAM will be developed and easily replicated across different states or even different jurisdictions within Wisconsin. Each locale brings with it unique idiosyncrasies in its data structure, in the inferences it hopes to make, the context of the assessment, and the surrounding policy environment.

Table 9 summarizes common ways data are used in educational evaluation and may help illustrate the distinction among the growth methods discussed in this paper. The group of models labeled "Snapshot Models" are the traditional AYP calculations, which rely on a single year of data (or two years in the case of Safe Harbor), but are standards-based in that they relate student scores to performance standards and do not report scores in relation to other students. The evaluations are based on different cohorts of students each year.

In contrast the group of models labeled "Growth Models" use longitudinal data. They follow the progress of individual students over time, and therefore require the ability to match student records for two or more tests. Some growth models support inferences about students or schools with respect to other students or schools included in the analysis. Others, such as growth-towards-the-standards models, evaluate whether a student is on track to reach proficiency at a certain point in time. Hence, the inference about a student is standards-based and not normative.

While some ideas from growth models may share common features (e.g., projections to a standard) the growth model itself and how it makes use of data will likely vary from state-to-state to reflect the differences in data and policy. This suggests that the most critical aspect of growth model development is for each constituency in Wisconsin to be able to determine with clarity what purpose their growth model is to serve. Because form follows function, consideration of the critical issues developed in Table 6 can be extremely useful starting points. With resolution of these issues, localities in Wisconsin can then begin work with model development and exploration with real data.

Table 9. Common student achievement models and their characteristics

Achievement Model (Section Number)	Examples (* indicates model for WI evaluation)	Purposes	Data and Resource Requirements	Cautions
		SNAPSHOT MODELS All require well-defined performance levels.		
Status (2.1)	• AYP	• Determine the proportion of students meeting performance standards • Current NCLB accountability	• Single set of test scores	• Cannot directly compare schools or districts due to unequal distribution of students • Does not provide information about growth
Improvement (not discussed in this paper)	• AYP Safe Harbor	• Determine change in the proportion of students meeting performance standards • Current NCLB accountability	• Two years of test scores	• Cannot directly compare schools or districts due to unequal distribution of students • Does not provide information about growth of individual students
		GROWTH MODELS All required longitudinal (linked) student records and at least two years of test scores.		
Simple* (3.1, 4.2.1)		• Determine growth of individual students	• Vertical scale	• Additional information is required to meaningfully interpret
Normative (3.1, 4.2.2)	• Similar Students Index* • Growth Percentiles	• Compare individual students' growth with other students' growth • Determine "normal" growth	• Well-defined reference groups	• High- and low-performing students may show anomalous growth patterns

Achievement Model (Section Number)	Examples (* indicates model for WI evaluation)	Purposes	Data and Resource Requirements	Cautions
Growth towards the Standards (2.3.1, 4.2.3)	• Yearly Target • Projection Probability of Proficiency*	• Determine which students are "on track" to reach proficiency • Approvable for current NCLB accountability	• Vertical scale is preferred	• Future student performance is not known with certainty
Value Table (2.3.2)	• Delaware model	• Recognize schools for the number of students who are proficient or who are making progress to proficiency • Approvable for current NCLB accountability	• Well-defined performance levels • Assignment of point values for different amounts of progress	• No objective way to assign point values for changing performance levels • Applies to school level and above
Value-Added Models (3.2, 4.2.4)	• EVAAS • RAND • Mixed-Effects*	• Determine the effect of a teacher, program, school, or district on student learning • Separates school effects from non-school effects Methods applicable to other growth models, by using covariates	• Specialized software • Significant statistical knowledge • Capacity for professional development on interpretation and use of results Include student characteristic	• No consensus on best method

The implementation of growth modeling and/or VAM is part of the wider movement toward Education Science, as we indicated at the beginning of this paper. Implementation of an innovative and effective reporting system that is designed to meet the needs of the school communities in Wisconsin is a challenging opportunity and one that will require a considerable effort by all the relevant constituencies including the Department of Public Instruction, policy analysts, administrators, teachers, and of course parents. Each needs to be a part of the discussion and each needs to come to the realization that asking the right questions, obtaining good data, analyzing insightfully, and carefully implementing interventions supported by the results will help everyone increase the level of success of the schooling process.

A commitment to the central purpose of schools helping young people achieve their fullest potential will be critical to the ultimate success of this effort. Recognizing that science has much to offer education – just as it has to medicine, engineering, and business – is critical to Wisconsin's success in this endeavor. While recognizing the ambitions of this effort it is also important to recognize its challenges. It is best to start with small steps while moving steadily forward in the direction determined by the various stakeholders in Wisconsin. The growth model evaluation project is a good first step.

References

Adcock, E. P. (1995). Value added effective schools study for elementary schools: 1994 Maryland school performance assessment program results. Research report no. 36-9-95. Maryland: Prince George's County Public Schools. Research, Evaluation and Accountability Office.

Amrein & Beardsley (2008) Methodological concerns about the education value-added assessment system. *Educational Researcher*, Vol. 37, No. 2, 65-75

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

Betebenner, D. W. (2008). Norm and Criterion-referenced student growth. *NCME*, New York, NY.

Betebenner, D. (2005). The importance of performance standards in measures of student growth. CCSSO conference on Large-Scale Assessment, San Antonio, TX.

Betebenner, D. & Shang, Y. (2007) Norm and criterion referenced student growth. CCSSO.

Briggs, D. C., Weeks, J. P., & Wiley, E. (2008). Vertical scaling in value-added models for student learning. National conference on value-added, Madison, Wisconsin.

Choi, K., Seltzer, M., Herman, J., & Yamashiro, K. (2007). Children left behind in AYP and non AYP schools: Using student progress and the distribution of student gains to validate AYP. *Educational Measurement: Issues and Practice,* fall, 21-32.

CCSSO (2007). Summary of state-proposed growth models. CCSSO.

CCSSO (2009) Guide to United States Department of Education Growth Model Pilot Program 2005-2008. CCSSO

Doran, H. C., & Cohen, J. (2005). The confounding effect of linking bias on gains estimated from value-added models. In R. Lissitz (Ed.), *Value-added models in education: Theory and applications*. Maple Grove, MN: JAM Press.

Doran, H. & Izumi, L. T. (2004). Putting education to the test: A value-added model for California. San Francisco, CA.

Doran H, Bates D, Bliese P, Dowling M (2007). "Estimating the Multilevel Rasch Model: With the lme4 Package." Journal of Statistical Software, 20(2). URL http://www.jstatsoft. org/v20/i02/.

Doran, H. (2008) On the topic of VAM: fixed and random effects models. The author can be contacted directly at AIR for a copy of this paper.

Doran, H., Lockwood, J. R. (2005). Fitting value-added models in R. *Journal of Educational and Behavioral Statistics*, Vol. 31 (2), 205-230.

USED (2007) State applications, decision letters and additional information. http://www.ed.gov/admins/lead/account/growthmodel/index.html This version was last updated on 1/08/09.

Hess, F. M. & Fullerton, J. (2009). Balanced Scorecards and Management Data, Harvard Center for Education Policy Research

Hill, R., Gong, B, Marion, S, DePascale, C., Dunn, J. and Simpson, M. A. (2006). Using value tables to explicitly value growth. In R. Lissitz (Ed.), *Longitudinal and value added models of student performance* (pp. 255-290). Maple Grove, MN: JAM Press.

Journal of Educational and Behavioral Statistics special issue on Value Added Modeling, Howard Wainer, Editor (2004).

Kane, T. & Steiger, D.O. (2002). Improving School Accountability Systems. NBER working paper No. W8156.

Kupermintz. H. (2003). Teacher effects and teacher effectiveness: A validity investigation of TVAAS (Tennessee Value Added Assessment System*). Educational Evaluation and Policy Analysis*, *25*, 287-298.

Linn, R. L. (2003) Accountability: Responsibility and Responsible Expectations. *Educational Research*.

Lissitz, R. W. (Editor, 2005). *Value Added Models in Education: Theory and Applications*, Maple Grove: JAM Press.

Lissitz, R. W. (Editor, 2006a). *Longitudinal and Value Added Modeling of Student Performance*, Maple Grove: JAM Press.

Lissitz, R. W., Fan, Wei Hua, Alban, T., Hislop, B., Strader, D., Wood, C., and Perakis, S. (2006b) The prediction of Performance on the Maryland High School Graduation Exam: Magnitude, Modeling and Reliability of Results. *NCME*, San Francisco.

Lockwood, J. R., McCaffrey, Mariono, and Setjodi (2007). Bayesian Methods for Scalable Multivariate Value-Added Assessment. *Journal of Educational and Behavioral Statistics,* Vol. 32, NO. 2, 125-150.

Lockwood, J. R., Doran, H. & McCaffrey, D. F.(2003). Using R for estimating longitudinal student achievement models. *The R Newsletter*, 3 (3), 17-23.

Managing for Results in America's Great City Schools: A Report of the Performance Measurement and Benchmarking Project (April, 2007).

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavorial Statistics*, vol. 31, No. 1, 35-62.

Martineau, J. A. (2007). Designing a valid and transparent progress-based value-added accountability model. Paper presented at the annual conference of the *American Educational Research Association*, Chicago, Illinois.

McCaffrey, D. F. and Hamilton, L. S. (2007) Value-Added Assessments in practice: Lessons from the Pennsylvania value-added assessment system pilot project. *Technical Report*, RAND Corporation.

McCaffrey, D. F., Koretz, D. M., Lockwood, J.R., Hamilton, L. S. (2004). *Evaluating value added models for teacher accountability*. MG-158-EDU, 154 pages, Pittsburgh: RAND Corporation.

Meyer, R. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, 16, 283-301.

Meyer, R. (2008). Value added model specification tests. National conference on value-added modeling, Madison, Wisconsin.

Meyer, J.H.F. and Sass, A.R. (1993). The impact of the first year on the learning behaviour of engineering students, *International Journal of Engineering Education*, 8, 328-35.

No Child Left Behind Act (2001). Public Law of PL 107-110. http://www.ed.gov/policy/elsec/leg/esea02/index.html

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Education and Behavioral statistics*, 103-116.

Sanders, W. L., Saxton, A. M., and Horn, S. P. (1997). The Tennessee value-added assessment system: A quantitative, outcome-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.

Sanders, W. L. (2006). Comparisons among various educational assessment value-added models. The Power of Two -- National Valued-Added Conference, Battelle for Kids, October 16.

Schafer, W. D. (2006). Growth scales as an alternative to vertical scales. *Practical Assessment Research and Evaluation*, 11(4). Available online: http://pareonline.asp?v=11&n=4.

Spellings (2005, Nov.). Secretary Spellings announces growth model pilot /Press Release/. U.S. Department of Education. (Retrieved August 7, 2006 from http://www.ed.gov/news/pressreleases/2005/11/1182005.html)

Thum, Y. M. (2001). Measuring progress towards a goal: estimating teacher productivity. *Sociological Methods and Research*, Vol. 32, 153-207.

Webster, W. J. (2005). The Dallas school-level accountability model: The marriage of status and value-added approaches. In Lissitz, R. W. (Ed.). *Value added models in education: Theory and applications*, Maple Grove: JAM Press.

Wisconsin Center for Education Research (2008). National conference on value-added modeling. http://www.wcer.wisc.edu/news/events/natConf_papers.php

Wright, S. P., Sanders, W. L., & Rivers, J. C. (2006). Measurement of Academic Growth of individual students toward variable and meaningful academic standards. In Lissitz, R. W. (Ed.). *Longitudinal and Value Added Modeling of Student Performance*, Maple Grove: JAM Press.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans and P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273-283). New York, Springer