Using Value Tables to Explicitly Value Student Growth

Paper presented at the Conference on Longitudinal Modeling of Student Achievement
November 8, 2005

Richard Hill, Brian Gong, Scott Marion, Charles DePascale
Jennifer Dunn, and Mary Ann Simpson

The Center for Assessment

Background

Many states are working to include a measure of school progress in their accountability systems. Some are planning to have this accountability system operate independently of, and in addition to, their accountability design in response to the requirements of the No Child Left Behind Act (NCLB). Others would like to include this element in their NCLB accountability designs, and are wondering what options the U.S. Department of Education (USED) will allow.

This is the context into which this paper is written. While none of the ideas in here have been accepted by the USED (and many haven't even yet been proposed), they are all generated with an intent to make them sufficiently compatible with the goals of NCLB so that they might be accepted as part of an entire NCLB accountability package.

Assessment and accountability vs. research and evaluation. As a starting point, we want to clarify the distinction between a data analysis system that is designed to estimate variance in scores due to teacher effectiveness for purposes of research and evaluation (a system that is a *passive describer* of "what is") versus one that is designed to be incorporated into an accountability system whose main objective is to improve current student performance (a system that is an *active element in changing* "what is"). The system we are proposing is far too unsophisticated to be considered a reasonable replacement for the models already in place for research and evaluation. But we will argue that those latter systems are far to complex to be used effectively in an accountability system, which is the intended use for the system we will be presenting in this paper.

Assessment vs. accountability. The second important point to note is that assessment and accountability are two different processes. Assessment is the process of acquiring, summarizing, and reporting information; accountability is the process of assigning consequences to the assessment results. Valid accountability requires valid assessment, although it certainly is possible to have a valid assessment that leads to invalid accountability. This paper will assume that valid assessment information exists; the goal now will be to create a set of rules that will lead to valid accountability decisions and encourage improved educational practice and student outcomes.

Validity of accountability. To be valid, an accountability system must, at a minimum, correctly sort accountable units into those that are meeting the stated goals of the system vs. those that are not. If, for example, an accountability system intends to reward schools where the

most effective instruction is going on, but then bases its decisions solely on whether sufficient students in the school are performing at the Proficient level or higher, the accountability system will lack validity. For such an accountability system to be valid, it would have to be immaterial to the school whether the entering students were all high performing or all low performing. Since most schools would find it substantially easier to get all their students to the Proficient level if they were already at that level before the school year began, it is clear that an accountability system based solely on measuring students' performance at the end of a given year cannot be validly measuring the effectiveness of the instruction in the school. If accountability systems that intend to reward effective instruction are going to be valid, some measure of student growth must be a major element included in the accountability design.

Goals of accountability. A final distinction between assessment and accountability comes with a goal of accountability—to change behavior. Accountability systems are put in place generally because there is a belief that changes in behavior by some of the participants in the system will lead to improved outcomes for the system as a whole. If that is the case, then accountability systems must be evaluated in terms of their likelihood to influence behavior in the intended direction. There are several elements that must be in place for this to happen:

1. Participants must understand what is expected of them—in advance
2. Participants must perceive that the goals are accomplishable
3. Participants must have the resources to accomplish the goals
4. Participants must perceive that accomplishment of the goals are within their control (at least partial control)
5. The incentives (i.e., rewards and consequences) must be consistent with the effort it will take to meet the goals

When educational accountability systems fail to accomplish their goals, it is usually because at least one of these elements is absent from the design. It may be assumed incorrectly, for example, that teachers have the necessary training and skills to accomplish the goals of the program, or that students are sufficiently motivated when taking the assessments to demonstrate fully the learning they have acquired. NCLB provides a good example of a goal that participants perceive will not be accomplished: the expectation that all students will be proficient by 2014. When participants perceive that the expected goal of an accountability system is unrealistic, they may fail to make even the incremental improvements that the advocates for educational improvement might have really intended.

Status vs. progress. At this point, it will be important to introduce terms that will be used throughout this paper. One way of categorizing accountability designs is whether they measure "status" or "progress" (or both). A design that measures status looks at the performance of students in a school at a given time, without regard to how students in that school scored previously. A design that measures progress looks at the performance of students in a school at a given time relative to how students in that school performed at a prior time.

"Progress" can be further delineated into "improvement" and "growth." Improvement measures change across cohorts, keeping grade constant, whereas growth measures change across grades, keeping the cohort constant. A design that looks at the performance of this year's

fourth grade students compared to last year's fourth grade students is using improvement as its criterion; a design that looks at the performance of this year's fourth graders compared to how those same students did last year (typically as third graders) is measuring growth. An advantage of an improvement design is that the outcome measures remain constant; an advantage of a growth design is that the students remain constant. If measures across two grades are highly coherent, a growth design will produce more reliable results. If measures across two grades have little coherence, it usually will be better to keep the outcome measure constant across years (an improvement design) than the students (a growth design).

Two additional points about progress measures need to be made. First, staff at the Center for Assessment feel that measures of both status and progress generally should be incorporated into accountability systems, and they should not be combined. Trying to obtain one total score for a school using both measures produces two schools with similar scores, but one with high status and low progress, while the other has low status with high progress. The public does not see these two schools as equivalent, and the entire accountability system loses credibility as a result. The two scores should be reported separately and not combined; when one overall judgment needs to be made about a school, it generally is better to require the school to have an acceptable score in one or both of the measures to be judged as an acceptable school overall.

Second, Braun (2005) noted that progress scores are really nothing more than conditional status scores. That is, any progress score is doing nothing more than answering the question, "How are you doing this year (status), given how you were doing last year (conditional status)?" While it is not necessary to view progress this way to measure it, it is a position that can help one resolve knotty questions; for example, why a growth model could be used when the tests across the two years aren't on the same scale—or for that matter, even measure different constructs. Braun's conception of the problem allows one to create valuable designs in these environments. Prior information about a student or a school is valuable so long as it correlates with current information. Tests do not need to be scaled across grade levels or be based on the same content standards for them to measure progress effectively, so long as there is a statistical relationship that can be established between the prior and the current information. While we would, of course, clearly desire to have prior information that is connected as directly as possible to the progress scores, it is not an "all or nothing" issue: accountability designs can be considerably improved by incorporating prior information that is well correlated with post-test scores, even if there is not a direct connection between the two scores.

Some Existing Measures of Growth

Several ways of measuring student growth already exist, which may lead one to legitimately question why yet another is being proposed in this paper. The answer is that the underlying philosophical consequences of these existing systems are usually not explicitly stated, but if they were, they would be counter to many of the principles of NCLB.

One major category of growth models is those that use regression and other techniques to compute an expected score for a student in one grade, given the student's performance in a previous grade or in previous grades. Hierarchical linear models (Raudenbush & Bryk, 2002; Thum, 2006) and the Tennessee Value Added Accountability System (TVAAS) (Sanders,

Wright, & Rivers, 2006) are examples of this class of growth models. While these models are excellent ways of describing "what is," and attempt to extract the impact of effective teaching from the data, they are not good models to use for accountability, and they are especially problematic when used for NCLB accountability.

The reasons why current regression models are not good models for accountability in general include the following:

1. They involve complex calculations, and the calculations are not made until the school year is completed. Thus, even though a school would (or at least, should) know the scores of its incoming students, it has little idea when testing is complete, even if it reasonably estimate the performance level each student had attained, whether it met the required level of achievement for its students. Thus, to the extent that an accountability system should establish clear goals and have participants working to accomplish those goals, these models are unsatisfactory. Schools that receive unsatisfactory scores long after the school year is over and the students are gone are likely to view such an accountability system as a "gotcha" rather than as a tool that encourages them to increase student learning. In a good accountability design, each teacher would know when the school year *began* what level of achievement would be required for each student at the end of the year, and thereby have a specific goal for each student.
2. In these systems, scaled scores are generally used as the criterion measure, in contrast to the performance levels that now accompany standards-based testing. Rather than keeping schools focused on what performance level students are achieving on the tests, these systems look at the amount of scaled score points of growth. As a result, the reporting of "growth" is likely to be completely disconnected from the reporting of "status." In addition, while it is reasonable to expect that a teacher should know at the end of the year whether a student is performing at, say, the Basic or Proficient level, but it is not reasonable for a teacher to know whether the students in the class have gained an average of, say, six scaled score points during the year.
3. Again, in contrast to the current focus on standards-based assessment, these models use a norms-referenced basis for measuring student growth. There is no standard for how much growth is "good enough," only whether growth is more or less than others, or more or less than growth in previous years. Schools never know whether they have done enough to improve student learning—only whether others have done more than they have. Standards for school improvement are based on "what is," or "what has been," rather than on "what should be."

These models are particularly inappropriate for use with an NCLB accountability design[1] for the following additional reasons:

[1] NCLB has, at its core, certain basic principles about national education priorities: in particular, assuring that all students who are not proficient in reading and mathematics to reach those levels. Other priorities might be established, and these different priorities would lead one to develop a different accountability system. However, most states right now are trying to ensure that their accountability system is compatible with NCLB requirements, and so we will continue to focus on that design in the initial stages of this paper.

1. They are compensatory. Growth by any student at any point along the scaled score distribution is deemed the equivalent of that amount of growth by any other student at any other point along the scale. Thirty points is thirty points, regardless of whether that growth came from a top-scoring student or one near the bottom. NCLB, in contrast, focuses on increased student achievement for students below the state's standard for proficiency. For NCLB accountability, schools cannot use additional progress for students who are already proficient to compensate for low levels of progress for students who are not yet proficient.
2. In a similar manner, there is no focus on whether students who are below proficient are moving closer to that standard as they progress from grade to grade. A school might be making little progress with its students below proficient, yet still be judged as highly effective if it has exceptionally high gains for its highest performing students.

In response to some of these concerns, some other measures of growth have been proposed. The "REACH" accountability system (Doran & Izumi, 2004) and the "Hybrid Success Model" (HSM) (Kingsbury, Olson, McCahon, & McCall, 2004) are two examples. In both these models, one calculates the distance that students below Proficient are from the Proficient target. Next, one uses that distance, and the number of years a student has remaining to reach the Proficient target, to calculate a target of growth for the student for the given year. One then assigns a value to each student that represents the percentage (REACH) or proportion (HSM) of the target achieved (that is, the current score minus the pre-test, with the difference divided by the target). Scores less than 1 are assigned to students who make less than the targeted amount of growth; students who achieve more than the targeted amount of growth earn a value greater than 1. The major difference between REACH and the HSM is that students at or above the Proficient level at the time of pre-testing are excluded from REACH; HSM assigns a constant growth target to these students and includes them in the calculations for the school. Note that one major difference between these two models and the regression models outlined earlier is that these require that calculations be done student by student, whereas the regression models simply compute averages for the school as a whole. REACH and HSM therefore maintain a focus at the student level that is more consistent with the intention of NCLB than the regression models.

Note also that while the authors of both the REACH and HSM models have conceptualized the use of these models with a vertical scale, only minor modifications in the design would be necessary to implement them without a vertical scale, so long as standards across grade levels were articulated. That is, suppose we knew that the standard deviations of scales at different grade levels were equal, and we knew that the standards for Proficient were each equivalent for their corresponding grades. Then, if a student was 15 points below Proficient at grade 3, and we wanted the student to be Proficient by the end of grade 6, we would know that the target for the student at the end of grade 4 was 10 points below Proficient, and we could assign a value equivalent to "proportion of gain achieved" by observing the student's score at the end of grade 4 relative to Proficient. If the student's score was within 10 points of Proficient, he/she would earn full credit or more for his/her school, while a score more than 10 points below Proficient at that grade would earn less than full credit.

While these two accountability designs are more consistent with the goals of NCLB, they still have several of the drawbacks associated with the other designs presented earlier. While considerably simpler than the earlier ones, the calculations associated with them still are fairly complex. More importantly, they use scaled scores (and the assumption that vertical scales can be validly created), diverting attention from the proficiency standards on which we want teachers to be focusing.

Introduction to Value Tables

The basic idea behind using a Value Table to create an accountability system is to look at the achievement level a student earns one year, compare it to the achievement level earned the previous year, and then assign a numerical value to that change. Higher values are assigned to results that are more highly valued.

Table 1 shows a rudimentary Value Table. This table has several flaws that will be discussed later in this paper, but it will serve as a beginning to illustrate the basic idea of a Value Table.

Table 1

A Rudimentary Value Table (Value Table A)

Year 1 Performance Level	Year 2 Performance Level			
	Below Basic	Basic	Proficient	Advanced
Below Basic	0	100	150	150
Basic	0	50	125	125
Proficient	0	0	100	100
Advanced	0	0	100	100

With this Value Table, students earn:

- 0 points for their school if they score at the Below Basic level in Year 2, or if they fall from Proficient or Advanced to Basic.
- 50 points for their school if they maintain performance at the Basic level
- 100 points for their school if they score at the Proficient or Advanced level two consecutive years.
- 100 points for their school if they move from Below Basic to Basic.
- More than 100 points if they move from below Proficient to the Proficient or Advanced level.

The score for the school is simply the average of the points earned by the students in the school. Suppose, for example, a school had two students, both scoring at the Basic level in Year 1. If one student continued to score at the Basic level in Year 2 testing and the other progressed to Proficient, then the school would get 50 points for the first student and 125 for the second, or an average of 87.5. Whether a score of 87.5 should be considered an acceptable score is an issue

we will discuss in a later section of this paper, but for now, let's suppose it is. Note the simplicity of the system. A school could, in fact, establish goals for each of these two students at the beginning of the year and know that if it met those goals, its score at the end of the year would be 87.5. The simplicity and transparency of this system, combined with its focus on performance levels instead of scaled scores, increase the likelihood that it would be an appropriate motivator. A school would know precisely what it needed to do to meet the goals required of it by the state. If a score of 100 was required, and the school had three Basic students, it would know, for example, that it needed to have two of them progress to the Proficient level by the end of the year, with the third student remaining at Basic. Then, the following year, with two students at Proficient and one at Basic, it would need to keep the two Proficient students at that level and raise the Basic student to Proficient if it wanted to achieve a score of 100 for a second consecutive year.

Issues to Consider in the Development of a Value Table

As we started to work with states to include the concept of a Value Table in their accountability system, we found that there were problems with some of the Value Tables that seemed, at first blush, to be obvious choices. The purpose of this section will be to describe what some of those problems were, and the steps we took to correct them.

We started working with Value Tables in a state that currently places each of its students on one of five performance levels, which we will refer to as Levels I-V. Students have to score at Level III or higher to be considered Proficient for purposes of NCLB. In the accountability system that the state was using prior to NCLB, students were assigned scores of 0, 50, 100, 150 and 200 for Level I through Level V, respectively, to calculate the status score for schools. As the state began to consider adding a measure of student growth to its accountability system, it seemed that a natural extension of this would be to use the Value Table provided in Table 2.

Table 2

A Value Table Initially Considered
(Value Table B)

Year 1 Performance Level	Year 2 Performance Level				
	I	II	III	IV	V
I	100	150	200	250	300
II	50	100	150	200	250
III	0	50	100	150	200
IV	-50	0	50	100	150
V	-100	-50	0	50	100

With Value Table B, a school earns 100 points for maintaining students at their previous performance level, gains an additional 50 points each time a student moves up a level, and loses 50 points each time a student moves down a level. This seemed to be the system that would

most closely reflect the values that were reflected in the existing status system—growth at all levels was valued equally.

The first realization that this Value Table would be unacceptable came by simply looking at what the results would have been for a previous year. We matched students across years at one pair of grades (successfully obtaining a prior year score for about 90 percent of the students tested in the second year) and computed results by performance level. In Table 3, we report the percentage of students at each performance level in Year 2, given the student's performance level in Year 1, as well as the average growth score earned by students at each performance level, using the values provided in Value Table B.

Table 3

Percentage of Students at Each Performance Level in Year 2,
Given Performance Level in Year 1

Year 1 Performance Level	Year 2 Performance Level					Average Growth Score
	I	II	III	IV	V	
I	64	27	8	0	0	120.5
II	24	43	32	1	0	105.0
III	4	18	64	13	1	94.5
IV	0	2	39	51	8	82.5
V	0	0	10	53	37	63.5

If one used Value Table B, students at the lowest levels would produce the highest scores for their school, while the higher-achieving students would produce lower scores. Indeed, the correlation between schools' status scores and the scores they would receive from this Value Table was -.23. One obvious problem with the system is that students at Level V can earn no score higher than 100 for their school, while students at Level I can earn no fewer than 100. Reflection on the problems with this table led us to a principle that we have had to return to time and time again: when policy makers think of these tables, they are thinking in terms of true scores, but the data we will have will be observed scores, subject to regression. The regression effect is obvious in Table 3. Students at Levels I and II tend to move to higher levels the next year, while students at Levels IV and V tend to move down.

After viewing our first proposed Value Table, state policy makers created two additional rules for us to follow: (1) No value should be less than zero, and (2) the value for any student that was at Level I in the second year should be zero. In addition, they told us that students that maintained performance at higher levels should be assigned more points than students that hold their own at lower levels. With that direction, we created the Value Table presented in Table 4.

Table 4

An Alternative Value Table, with Observed Results for Each
Year 1 Performance Level when Applied to Actual Statewide Data across Years
(Value Table C)

Year 1 Performance Level	Year 2 Performance Level					Average Score
	I	II	III	IV	V	
I	0	200	250	300	230	74.0
II	0	100	130	180	230	86.4
III	0	50	100	150	200	94.5
IV	0	20	70	120	180	103.3
V	0	0	40	100	160	116.2

Value Table C overshoots the mark when one takes regression into account. Students who start at lower performance levels tend to earn the fewest points, while students at the higher levels earn more. Whereas the correlation between growth score and starting status scores was -.23 for Value Table A, it was .61 for this Value Table.

Thus, our next thought was to create a Value Table that we could call "neutral;" that is, one that took regression into account and produced average scores for each Year 1 performance level that were roughly equal. Then, we would ask policy makers to tweak that neutral table, rather than creating one from scratch that we knew would not take regression into account. There are two sources of regression that one might consider. The first is due to measurement error within year, and the second is the regression of students across years. The first source certainly should be taken into account, and can be computed readily if one knows the reliability of the tests being used. The second source is questionable. We know that, even if we had true scores, some students would move from, say, Level II to Level III from one year to the next even if they had average instruction—it just would have been their year to grow. Students don't grow evenly every year, even if the instruction they have is constant from year to year. Some will have a banner year in, say, grade 3, and then perhaps show little growth from grade 3 to grade 4, while other students will show the opposite pattern. Thus, there will be some "churn" even if instructional effectiveness is constant for all students, and it likely will occur to a greater extent than one might think, since an underestimate of a student's performance level one year automatically introduces error into the measurement of gain. Doran and Cohen (2005) showed that linking error could lead to significantly greater uncertainty in the measurement of growth than researchers were used to seeing. On the other hand, looking at the actual results of how students change levels from one year to the next almost certainly overstates the amount of true regression going on. There certainly are some students who grow from Level II to Level III because they had truly superior instruction during that year, not because of regression effects. So if one applied statewide observed results to a truly neutral Value Table, one should expect somewhat higher scores for students at the lowest levels and lower scores for students at the highest levels. We had not considered that at the time we started this work, so we did not take that into account when we proposed a third Value Table for the state. At that time, we thought it was neutral; now, we would question whether it over-corrects for regression.

However, given our understanding of the regression issue at that time, we proposed the Value Table presented in Table 5.

Table 5

Another Alternative Value Table, with Observed Results for Each Year 1 Performance Level when Applied to Actual Statewide Data across Years (Value Table D)

Year 1 Performance Level	Year 2 Performance Level					Average Score
	I	II	III	IV	V	
I	0	200	400	500	600	86.0
II	0	100	150	200	250	93.0
III	0	50	100	150	200	94.5
IV	0	10	60	110	160	92.5
V	0	0	20	90	120	94.1

This table produced results with which the policy makers in the state felt comfortable. Although we almost certainly have overstated the amount of regression in our calculations of the "average score" for each Year 1 Performance Level, and thereby understated the difficulty teachers with lower performing students will have achieving the same scores from this Value Table as teachers with higher performing students, policy makers were satisfied with that. One of the considerations that led to their acceptance of this Value Table is that there is a perception within the state that lower performing students probably already are receiving less effective instruction, and the fact that their scores tended to be lower was probably a reflection of that fact.

Note that Value Table D is not completely aligned with NCLB goals. The policy makers did not want the Value Table to completely ignore NCLB goals, but at the same time, they had objectives for state growth that were more expansive than those of NCLB. For example, the state accountability system includes tests in content areas beyond reading and mathematics, and they do have concerns about the progress that students make even when they are at Level III or above. As a result, they wanted to distinguish among differences in the three levels at or above Proficient (Level III), and felt that it was in the best interest of the state to do so. As a result, there is some degree of compensation in the system; a school's growth score can be raised by having higher level students improve more. Note, however, the severe penalty for having any student perform at Level I, along with the high score assigned for moving students from Level I to Level II, as well as the relatively low score assigned for keeping students at Level II. So while the final table is compensatory to some degree, it places great pressure on schools to move students to Level III (proficient) as quickly as possible; to the extent it does that, it is compatible with the stated goals of NCLB.

In the process of generating this table, we looked at some other statistics that were of interest. We reported earlier that the correlation between school's status score and its growth

score, using the Value Table B, was -.23, and for Value Table C, it was.61. That same statistic for Value Table D was.44. We also did a reliability study by drawing students with replication from the data base[2]. We found that the reliability of schools' mean status scores was .99. The reliability of the growth scores calculated using Value Table B was .92; for Value Table D, it was .94. This is in contrast to the reliability of the existing system for measuring school progress, which was to compare the school's status score in Year 2 with its status score in Year 1 (an improvement model), which was .87. Thus, one very positive result for Value Tables was an improvement in reliability over the system currently in use.

As we began to discuss this approach to measuring student growth with other states, a recurring theme was the desire to measure low-performing students' progress toward proficiency. In response to those concerns, some have decided to subdivide the performance levels below Proficient (usually two) into two or three. Thus, it becomes possible to measure a student's growth from, say, a low Level I to a high Level I. With four or more levels between the bottom of the scale and Proficient, it becomes possible to measure finer degrees of student progress. Note that the reason for making these subdivisions is not to improve reliability; at the school level, where most of the decisions are being made, dividing the scale into finer pieces has a minimal impact on reliability. The rationale is to be able to reward smaller steps toward a successful end for students. As a result, it is important to establish that each level is distinguishable from its neighboring level. One state's review committee has offered the guidance that the width of the intervals should be, at a minimum, something greater than the measurement error for the test, and that it should be possible to describe differences between the achievement of students in adjacent performance levels, just as the current performance descriptors do.

Another concern is how schools get rewarded for students who move up or down and then back to their starting position. For example, suppose a student starts at Level II in one grade, moves up to Level III in the next grade, and then back down to Level II again the year after that. If we use Value Table D, that student would get 200 points after the second year and 60 points after the third year, or a total of 260 points across the two years. A student who stayed at Level II all three years in a row would earn 100 points each year, or a total of 200 points across the two years. It is questionable whether the final outcome for the first student is any better than that for the second, and if it is not, the validity of the Value Table is undermined.

An additional issue to pay attention to is the assignment of points to the extreme off-diagonal cells. It is unlikely that a student would truly change from Level I to Level V in a year; similarly, it likely is untrue that a Level V student one year would decline to Level I the following year. When assigning points to these cells, one has to ask whether a change from Level I to Level V can be considered any better than a change from Level I to Level III. A change from Level I to Level V would more likely be due to a student's changed motivation in taking the test the second year than it would be that extreme a change in the student's true achievement level. Thus, it is quite questionable whether a change from Level I to Level V should be worth 500 points when a change from Level I to Level III is worth 300 points.

---

[2] More details about procedures for determining the reliability of school scores are available on the NCIEA website (Hill, Hill and DePascale). The specifics of the values are not the primary point in this section; what is important here are the relative differences among the reliabilities of status scores, improvement scores and growth scores.

It is clear that this system of Value Tables, regardless of which table is finally used, has real advantages over other measures of growth currently being advocated. It clearly is simpler to understand and implement, and it maintains a focus on the performance levels with which we hope teachers will become familiar. It also makes more explicit what kinds of outcomes are most valued, given the philosophy of policy makers. Value Tables take the discussion of kinds of improvement that are expected from "Any gain by any student is valued" to "These kinds of gains by these kinds of students are what we value the most."

However, it is of interest to know what the relationships are between this method of measuring growth and others that are being proposed. For this reason, staff at the Center for Assessment took data from one state and calculated schools' growth scores using each of three Value Tables, and compared those results to two other more traditional ways of computing student growth. The first alternative was a two-level analysis of covariance, using students' first year test scores as the covariate ("ANCOVA"), in which each school's score was expressed as the deviation from the overall predicted status. The second was a two-level hierarchical linear model that estimated slope parameters for schools, expressed as a deviation from the average slope for students statewide ("HLM slope"). Further details on these analyses are provided in Appendix A. The correlations of school scores among several statistics are reported in Table 6.

Table 6

Correlations among Several Measures of School Growth

	ANCOVA	HLM Slope	Value Table B	Value Table C	Value Table D
Year 1 Status	.70	-.19	-.20	.65	.44
Year 2 Status	.88	.12	.08	.82	.64
ANCOVA		.57	.56	.93	.85
HLM Slope			.98	.53	.67
Value Table B				.54	.69
Value Table C					.95

First of all, the correlations show that it matters which Value Table you choose, so decisions about what values to insert into the table should not be taken lightly. Further, the correlations show that Value Table B provides essentially the same information as HLM slope, while Value Table C is a fairly close match to ANCOVA. Interestingly, when policy makers had an opportunity to define what they truly wanted in an accountability system, they favored Value Table D over the other two. That result, in turn, would imply that if policy makers truly understood what results statisticians were providing in their "growth" analyses, they might not find them as acceptable as they think they do. Note that ANCOVA results were well correlated with the Year 1 status scores, meaning that the schools that have high-achieving students are likely to continue to be judged successful if this method of assessing student growth is chosen. HLM slope results, on the other hand, are negatively correlated with schools' starting positions.

System Requirements

For a system like this to work, there are a few basic requirements:

1. You must have annual testing at some consecutive grade levels.
2. You must have the ability to track students across years, so that you can match results student by student.
3. You must have articulated standards, so that the meaning of each performance level has a consistent meaning across grades.

The first two requirements are fairly easy to meet in the current testing environment. NCLB requires all states to have testing for every student at every grade between 3 and 8 every year, so the annual testing requirement is a given. Similarly, most states either have or are developing systems for tracking their students across years, again, in response to NCLB requirements. States *should* have articulated standards across grades, although many have not yet gone through the process of establishing the relationship of their standards across grades. When testing was done at limited grades, there was not the same concern about whether standards had consistent meaning across grades. But now that testing will be done at several consecutive grade levels, the articulation of standards across grades takes on a new level of importance. Parents will be tracking their children's performance levels across years (even if the state's accountability system does not!) and certainly will be expecting explanations if their child moves from Proficient at the end of grade 3 to Basic at the end of grade 4—and a response of "the standards at the two different grades can't be interpreted relative to each other" certainly won't be considered acceptable. So the first two requirements already are either in place or will be in place shortly in most states, and the third requirement will have to be met by states in order to reasonably report their results across years.

Using Policy Makers to Set Values

Once the requirements in the previous section are met, the final step is to get policy makers to concretely define the outcomes that are valued. Do they value a student moving from Proficient to Advanced? More or less than a student moving from Basic to Proficient? These kinds of questions need to be answered before the cells in the Value Table can be filled. In this section, we will present an outline for doing this that we have tried with some success in two states.

While the purpose of using Value Tables was to specifically incorporate the values of policy makers into the accountability system, we found it difficult initially to obtain feedback from that in a manner that we could use to revise the neutral Value Table. Discussions with policy makers would be vague; they could not express what they wanted valued more and what they wanted valued less. We felt we needed more specific information from them; for one thing, it is difficult to judge the validity of a particular system if one does not know what outcomes are most valued. Note, however, that there is another side to this issue—it often is easier to get agreement on something if one is vague about what the agreement is. However, we wanted more explicit statements about what was being valued, so we created a process to force policy makers to express some of this information.

The first step was to create sets of cards, with 25 cards in each set. Each card represented one of the 25 cells in the Value Table, and was labeled with that combination. We divided the policy makers into teams of three or four and gave each team a set of cards, with instructions to order the cards from the most desirable outcome to the least. There were many parallels between how we ran this session and how a typical standard-setting session would be run. For example, after the teams had done their initial ordering, we had each explain the ordering they had created and the justification behind it. Following that discussion, we had them review their decisions in the first round and make any changes they thought appropriate, given the input from the other groups. After the session was over, we collected information on each group's orderings and computed the mean rank for each cell.

The session did indeed accomplish one major objective, in that it got the policy makers to be much more specific about what they valued. A strategy that one group employed (which was then immediately copied by almost all the other groups) was to look at each of the diagonals separately. That is, they first looked at the five combinations that had students maintaining their status from the previous year. As would be expected, they rated students maintaining status at Level V, then Level IV, and so on down to Level I. The next set of cards had the students all moving up one level from the previous year. In this case, moving from Level I to Level II generally was rated as a higher achievement than moving from Level IV to Level V. But once that was done, the committees had to wrestle with more specific questions, such as, was it more desirable to move from Level I to Level II than it was to maintain status at Level V across years? The discussions around these choices helped policy makers articulate, and us to understand, what cells they truly valued over others.

While the session went well, and was a process we generally would recommend, it did not go completely smoothly. One question that arose early in the process (and created an important point for the policy makers to reflect upon) was whether higher values should be assigned to those cells that were harder for teachers to accomplish or those that reflected how desirable the outcome was to the policy makers. That was an important point of clarification that might not have arisen otherwise.

The biggest problem we encountered, however (and did not realize until after the session was over), was that the policy makers were thinking of true scores and not observed scores. That is, when they were thinking of the relative value of a student maintaining status at Level III vs. Level V, for example, they were thinking in terms of the student's true level, not the observed result on the test. Given that students at Level III are scoring around the state average while students at Level V are scoring well above it, there is significantly more negative regression in Level V scores than there are for Level III scores. Therefore, if one wanted to assign the same value to students staying at Level III to those staying at Level V, the "V-V" cell would have to be assigned a far higher value than the "III-III" cell.

However, our thinking on this issue has not yet progressed much past this level. It is clear that policy makers will be thinking in terms of true scores, and we must develop the Value Table knowing that it must apply to observed scores, so we cannot directly translate their values

into the cell entries for the Value Tables. It is not clear how this should be done operationally, so this is one area in which future development is required.

Setting Goals

Once we have chosen the Value Table that will allow us to calculate a score for each student, and therefore each school, the next step would be to decide what score a school needs to attain in order to be judged successful or unsuccessful. Of course, one way to do that would be to simply compute the scores for all the schools in a state and choose a cut by a norm-referenced procedure. But since we are working in a standards-based environment and have selected the values in the Value Table to have some meaning, it seems more appropriate to establish a standards-based criterion for success.

If the philosophy underlying the development of the Value Table was that all students should be on track to becoming proficient, it may very well be possible to determine directly what score is required for a school to be judged successful. In such a Value Table, the values assigned to students who are above proficient in the baseline year and continue to be above proficient in the post-test year are assigned some value—say, something a little above 100—while students who are below proficient but making progress toward proficient receive values well above 100, and those who are below proficient and not making progress toward proficient receive values below 100. Table 7 shows such a Value Table. Note that in this Value Table, students at Levels III and IV are considered Proficient; Levels I and II are steps below Proficient.

Table 7

A Value Table that Might Be Used to Measure Progress toward Proficient (Level III)
(Value Table E)

Year 1 Performance Level	Year 2 Performance Level			
	I	II	III	IV
I	0	120	160	200
II	0	80	140	160
III	0	40	120	140
IV	0	0	100	120

Note that Value Table E would be significantly improved by subdividing Levels I and II, as noted in an earlier suggestion, so that we could make more fine-grained decisions about whether students are making progress toward Proficient. However, to keep the illustration as simple as possible, we will use Value Table E as written. With this Value Table, the goal would be for every school to have a score of 100 or more. Scores of 100 or more are assigned to students who are making progress toward Level III or are remaining there, while students who are not making progress toward that goal are assigned values of less than 100.

Another possible approach is to compute the "transition matrix" that meets the stated goals of the state. Betebenner (2005) has outlined this approach in more detail than will be provided here, but the following is the essence of his ideas on this subject. A transition matrix simply provides the conditional probabilities that students will move from one performance level to another over a period of time (usually a year). Table 7 was the observed transition matrix for one state over one year. If one multiplies the vector of the starting observed distribution of students across performance levels (the "initial state") by the transition matrix, the product is the proportion of students that will be in each performance level at the end of the year (the "final state"). For example (using tables provided by Betebenner (2005, p. 9)), students are placed into one of four performance levels, with 25 percent of the students in each level in the initial state. When the year is up, the final state is different from the initial state; 55 percent of the students are in the top two levels, compared to 50 percent at the beginning.

$$\begin{bmatrix}0.25\\0.25\\0.25\\0.25\end{bmatrix}^T \begin{bmatrix}.66 & .34 & .00 & .00\\ .10 & .59 & .31 & .00\\ .00 & .09 & .82 & .09\\ .00 & .00 & .36 & .64\end{bmatrix} = \begin{bmatrix}0.19\\0.26\\0.37\\0.18\end{bmatrix}^T$$

Now, taking that example one step further, suppose that same transition matrix was applied to the final state obtained in the previous paragraph—which, of course, becomes the initial state for the next year. That would provide the final state after two years. Suppose further that the initial state was observed at the end of grade 3, and the transition matrix held for seven consecutive years, until the students had completed grade 10. If we multiply the transition matrix seven times to the original starting distribution, we get the following result:

$$\begin{bmatrix}0.25\\0.25\\0.25\\0.25\end{bmatrix}^T \begin{bmatrix}.66 & .34 & .00 & .00\\ .10 & .59 & .31 & .00\\ .00 & .09 & .82 & .09\\ .00 & .00 & .36 & .64\end{bmatrix}^7 = \begin{bmatrix}0.075\\0.204\\0.582\\0.138\end{bmatrix}^T$$

Suppose a state had decided to use Value Table E, and came to the conclusion that the percentage of students at Levels III and IV in the final state was satisfactory (i.e., consistent with the state's long term goals). Then a school that is "on track" to meeting the state's goals for improved student growth would receive a score of .25 * (.66*0 + .34*120) + .25 * (.10*0 + .59*80 + .31*140) + .25 * (.09*40 + .82*120 + .09*140) + .25 * (.36*100 + .64*120), or 89.7. Thus, this state might set 90 as the required score for schools to be considered satisfactory.

Note that the more logical way to apply this method would be to establish the desired final state, the number of years to achieve it, and then knowing the initial state, compute the transition matrix that will meet the state's needs. There is no unique solution to this problem; there are many transition matrices will produce the required final state. However, one could compute a limited number of them, determine what scores would be obtained by schools producing such transition matrices, and establish a reasonable required score for schools from that analysis.

A Side Note: The Relationship between Value Tables and NCLB Calculations

As a side note, it is worth pointing out that NCLB accountability designs can be expressed as Value Tables. The "status" calculation for NCLB is the percentage of students at the Proficient level or higher. Table 8 shows how the status calculation of NCLB could be expressed as a Value Table.

Table 8

A Value Table Consistent with the
NCLB Status Calculation

Year 1 Performance Level	Year 2 Performance Level			
	Below Basic	Basic	Proficient	Advanced
Below Basic	0	0	100	100
Basic	0	0	100	100
Proficient	0	0	100	100
Advanced	0	0	100	100

In this case, the performance of students in Year 1 is ignored. If students score at the Proficient or Advanced level in Year 2, regardless of how they scored in Year 1, they earn 100 points for their school; if they score below Proficient, they earn 0 points. The average of all those scores will be the percentage of students Proficient or Advanced at the end of Year 2, independent of any performance in Year 1.

A Value Table can also be used to express the "safe harbor" calculation for NCLB. To achieve safe harbor, a school must reduce the percentage of non-Proficient students by 10 percent from one year to the next. Table 9 is a Value Table that reproduces those calculations. If a school's growth score is calculated through this Value Table, it will get exactly the same score as computing the percentage reduction of non-Proficient students. Thus, if a school's score using this Value Table is 10 or more, it makes Safe Harbor, and if the score is less than 10, it fails to make Safe Harbor. Note that Safe Harbor is typically computed as a result across cohorts, but there is nothing in NCLB that would prohibit it from being computed as a result within a cohort across years. NCLB says that a group makes Safe Harbor if "the percentage of students in that group who did not meet or exceed the proficient level…for that year decreased by 10 percent of that percentage from the preceding school year…" NCLB doesn't say whether the comparison from the previous year should be based on other students from that same group, or on the same students. Since NCLB had to be implemented before many states had consistent, standards-based testing at several consecutive grades, there was no choice but to use students from a different cohort in the previous year. But as states bring testing at all grades between 3 and 8 on-line no later than 2006, they might decide that the fairer comparison is within cohorts across grades. If that is done, then this example of a possible Value Table is completely appropriate.

Table 9

A Value Table Consistent with the
NCLB Safe Harbor Calculation

Year 1 Performance Level	Year 2 Performance Level			
	Below Basic	Basic	Proficient	Advanced
Below Basic	0	0	100	100
Basic	0	0	100	100
Proficient	-90	-90	10	10
Advanced	-90	-90	10	10

The entries in the cells in this Value Table might seem strange, so let's take a closer look at the calculations that would result. Suppose a school has 100 students, with 50 of them below Proficient in the first year. To meet Safe Harbor, the school would need to have 55 of their students at the Proficient level by the end of Year 2 (since 50 percent of the students are below Proficient in Year 1, and the school therefore must reduce that by 5 percentage points—10 percent of 50 percent). Suppose the school accomplishes that. A possible cross tabulation of performance across the two years for this school is shown in Table 10.

Table 10

A Possible Result for a School that Makes Safe Harbor

Year 1 Performance Level	Year 2 Performance Level	
	Below Proficient	Proficient or Above
Below Proficient	45	5
Proficient or Above	0	50

The school has 50 of its students below Proficient in Year 1 and 45 in Year 2. Its score, given the Value Table in Table 9, would be (45*0 + 5*100 + 50*10)/100, or 10. It exactly makes Safe Harbor.

Now let's extend the example by supposing that the school exactly makes Safe Harbor, but does so in a somewhat different way. In this example, some students who were Proficient in Year 1 drop below that cut score in Year 2, but the slide by some students is compensated by having more students who were below Proficient move above that standard. An example of this is provided in Table 11.

Table 11

A Second Possible Result for a School that Makes Safe Harbor

Year 1 Performance Level	Year 2 Performance Level	
	Below Proficient	Proficient or Above
Below Proficient	40	10
Proficient or Above	5	45

Just as with the first school, this school has 50 percent of its students below Proficient in Year 1, and reduces that to 45 percent in Year 2. Its score, given the Value Table in Table 9, would be (40*0 + 10*100 -5*90 + 45*10)/100, or 10. No matter what the combination created, a school's score using this Value Table will equal its reduced percentage of non-Proficient students, so a score of 10 or higher is a satisfactory result for a school and a score less than 10 is not satisfactory.

Areas for Further Research

Since we are just beginning to use Value Tables in state accountability systems, we have many areas of unknowns. In this section, we will focus on just two of them.

The first question came about from looking at the results for one state separately for whites and African-Americans. Tables 12 and 13 provide the initial states for both groups, as well as the transition matrix that was observed in the state for one pair of grades.

Table 12

Initial State and Transition Matrix for Whites

Year 1 Performance Level	Percentage of Students Initially	Year 2 Performance Level				
		I	II	III	IV	V
I	8	0.46	0.34	0.19	0.01	0.00
II	14	0.18	0.38	0.42	0.02	0.00
III	49	0.03	0.15	0.65	0.16	0.01
IV	23	0.00	0.01	0.39	0.49	0.11
V	6	0.00	0.00	0.10	0.50	0.39

Table 13

Initial State and Transition Matrix for African-Americans

Year 1 Performance Level	Percentage of Students Initially	Year 2 Performance Level				
		I	II	III	IV	V
I	31	0.59	0.29	0.12	0.00	0.00
II	26	0.28	0.41	0.30	0.01	0.00
III	37	0.08	0.24	0.60	0.08	0.00
IV	6	0.01	0.05	0.50	0.39	0.05
V	1	0.00	0.02	0.17	0.57	0.25

From the tables, it is clear to see that for each starting performance level, whites have a higher probability of being in a higher level at the end of Year 2 than do African-Americans. For example, only 46 percent of the white students who were at Level I the first year repeated at that level the second year, while 59 percent of the African-American students remained at that level. Similarly, white students who scored at Level IV or V the first year were more likely to stay at those higher levels the second year than were African-American students. Indeed, when we applied Value Table D to those data, whites scored higher than African-Americans at every starting performance level. The average growth score earned by white students was 97.3, while for African-Americans it was 79.5. Given that most observers within this state believe that white students get better teachers than do African-American students, we believed the results we were obtaining were consistent with what one might reasonably expect. However, we were surprised to see that the final state for whites (obtained by multiplying the transpose of the initial state by the transition matrix) was the same as the initial state, and the same thing was true for African-Americans. That is, despite the seemingly wide gulf between the two transition matrices, each was providing stasis for its respective group. This observation leads to an obvious, immediate question: If the gap between African-Americans and whites is remaining constant, why should white students receive higher growth scores? Clearly the answer has something to do with regression. Whites are regressing to the average for whites, while African-Americans are regressing the average for their group, and the former average is higher than the latter. This is an additional indication that we do not yet have a sufficient handle on how regression should be taken into account in the design of an appropriate Value Table.

The second issue is related then to the first. Suppose two schools have different initial states, and therefore have to make different rates of progress to meet a state-established long-term goal (for example, having 95 percent of the students proficient by 2014). Suppose further that the transition matrix for both schools show them to be on track to meeting that goal. Should both schools get the same score in a system that measures and values growth, or should the school with the more distant starting point receive the higher score? If the answer is that the school that has made more progress should get the higher score (which only seems fair), how do we interpret growth scores relative to the states' goal?

Summary

In this paper, we have proposed a method for measuring student growth that has several advantages and disadvantages when compared to other systems currently in use. While it lacks the sophistication and precision of those other methods, that might very well be one of its advantages. Because the calculations associated with it are so simple, schools would be encouraged to establish advance goals for their students that are clear and can be readily communicated in terms of content and performance standards, and know the consequences will be if they meet those goals. We believe such a system can be effectively used within an overall program designed to improve schools' educational programs.

Another advantage of this system is that it provides an opportunity for policy makers to have their values explicitly included in the accountability system. In this paper, we showed a Value Table (Value Table B) that proved to be wholly unacceptable to policy makers once they understood what the values in the table (and the consequences of the values) were. Yet this table produced the same correlation between status and growth scores that many of the current systems for measuring student growth do (a slightly negative correlation). An obviously realistic possibility is that policy makers would seriously question the values implicit in these other measures of student growth if they knew what those values (and their consequences) were.

On the other hand, we still have a long way to go to demonstrate that Value Tables produce measures of growth that truly mean what they seem to mean at first blush. Clearly, the issue of regression needs to be more clearly understood. What sources of regression should be taken into account, and which should not, when developing a neutral Value Table? What are the impacts of students' initial states on the distribution of school mean growth scores even when the Value Table remains constant? How do we translate policy makers' values, given to us on a true-score basis, into a Value Table that is fair when applied to students' observed results?

Finally, we need to understand better what the measurement costs are of this simple system, and whether adding some complexity would significantly improve the measurement of growth. Some existing systems use three to five years of previous data to establish a target for each student. While using this additional data clearly significantly improves the ability to predict a score for a student, it is not clear that the advantages remain, at least to that extent, when scores are aggregated to the school level. We have found, for example, that the reliability of school-level status scores are only marginally affected by increases in student-level reliability (Hill & DePascale, 2002). It is quite possible that the same situation holds true here, and increased complexity in the analysis system does not produce concomitant increases in the accuracy of the measurement of school gains. If that is true, then a strong case can be made for this simpler system; if not, then policy makers would need to know the trade-offs they would be making if they opted for a system that is simpler but has larger errors than one that is more complex.

Appendix A

Explanation of HLM analyses

Two different models were fit to student level datasets for ELA and Mathematics for the 2002-2003 datasets and the 2003-2004 datasets. Both models were two-level models (student and school) with random intercepts. The first model was an ANCOVA using the previous year's score as a predictor of current status. The second model was a slope model assessing change from year 1 to year 2. The slope was included as a random effect. The school code associated with any student represented the last school attended by a student in our data, i.e., a "school context" effect.

The "solution" option in the "random" statement in SAS PROC MIXED will output each school's random effect estimates for a given model. In the ANCOVA model, each school's intercept estimate represents its deviation from its predicted status. This status is predicted from the same prediction function for all schools. In the slope model, each school's slope estimate represents its deviation from the common predicted slope. All estimates are shrunken toward the grand mean (See p. 139 of Littell, Milliken, Stroup, & Wolfinger, 1996 for details).

Table 1 shows the fixed effects estimates for the two models' results with both subject areas and datasets. Note that in the ANCOVA model the term "slope" refers to predicted status and in the slope model, it refers to change from year 1 to year 2.

Table 1

Fixed Effects for HLM models for ELA and Math, 2002-2003, 2003-2004

	Intercept	Slope
Mathematics		
ANCOVA model		
2002-2003	27.90	.6498
2003-2004	27.95	.6430
Slope model		
2002-2003	71.24	2.49
2003-2004	73.34	1.27
ELA		
ANCOVA model		
2002-2003	30.52	.6400
2003-2004	31.40	.6356
Slope model		
2002-2003	77.30	2.24
2003-2004	78.98	2.05

References

Betebenner, D. (2005, June). *Performance standards as measures of educational effectiveness.* Paper presented at the annual National Conference on Large-Scale Assessment, San Antonio, TX.

Braun, H. (2005). Value-added modeling: What does due diligence require? In R. W. Lissitz (Ed.), *Value added models in education: Theory and Applications.* Maple Grove, MN: JAM

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

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

Hill, R. & DePascale, C. (2002). *Determining the reliability of school scores*. Retrieved November 22, 2005, from http://www.nciea.org/cgi-bin/pubspage.cgi

Kingsbury, G. G., Olson, A., McCahon, D., & McCall, M. S. (2004). *Adequate yearly progress using the hybrid success model: A suggested improvement to No Child Left Behind.* Retrieved November 22, 2005, from http://www.nwea.org/research/grd.asp

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models.* Cary, NC: SAS Institute.

Thum, Y. M. (2006). Measuring and comparing academic progress towards a standard using Bayesian performance profiles. In R. W. Lissitz (Ed.), *Longitudinal and value added modeling of student performance.* Maple Grove, MN: JAM

Sanders, W., Wright, S. P., & Rivers, J. C. (2005) Measurement of academic growth of individual students toward variable and meaningful academic standards. In R. W. Lissitz

(Ed.), *Longitudinal and value added modeling of student performance.* Maple Grove, MN: JAM

Raudenbush, S. W. & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed). Thousand Oaks, CA: Sage.