# Using Value Tables for a School-Level Accountability System

Paper presented at the NCME Annual Conference
April, 2006

Richard Hill
The National Center for the Improvement of Educational Assessment

## Background

The goal of this activity was to create a statewide school-level accountability system that judged schools on the basis of changes in student scores across years. In order to make the system effective (i.e., to create a system that changed school-level behavior in desired ways), it was felt that the system should have the following characteristics:

1. It would be simple to understand. Schools cannot fairly be held accountability to a system that they do not understand.
2. It would be computationally simple. State policy-makers felt that schools would make positive changes only if they were able to compute the results to which they were being held accountable. Under the desired system, a school would be able calculate in advance what its results would be, in contrast to other systems that might require weeks or months before the school found out whether its students had grown sufficiently by the end of the school year.
3. It would use the performance levels currently used in the state's assessment system rather than scaled scores. The state had chosen performance levels as its primary reporting statistic because they wanted those scores to develop meaning to school-level personnel over time. They wanted the accountability system to reinforce the use of these statistics, and therefore wanted them to be at the center of it.

Systems such as hierarchical linear models were rejected for this system because they fail all three criteria. While such systems are straightforward enough on their surface, it is hard for most people other than the most devoted data analysts to understand precisely how the models work. The computations involved are complex, requiring sophisticated software and computer far beyond the range of most schools. Finally, they employ scaled scores. We decided to create a different method of measuring school growth.

## Developing a Value Table

As an initial step was to consider an accountability system that compared the achievement level a student earns one year to that student's achievement level the previous year, and then assign a numerical value to that change. Higher values would be assigned to results that are more highly valued. We decided to called this matrix of assigned values a Value Table.

Table 1 shows the Value Table we first considered for the state's accountability system. Note that the state uses five performance levels with which to report its results. The levels are ordered from lowest to highest, with the third level being considered "proficient" for purposes of No Child Left Behind (NCLB) accountability. This table seemed fairly straightforward to develop. When students maintained their performance from one year to the next, the school

earned 100 points; the school earned 50 additional points each time a student went up one performance level, and lost 50 points for each performance level that the student went down.

Table 1

A Value Table Initially Considered
(Value Table A)

| 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 |

However, when we began to apply data to this Value Table, it was clear that it had some significant deficiencies. It was most obvious when we looked at the current changes in student performance across years. Table 2 provides the percentages of students who performed at each performance level in 2003, given the students' result in 2002. The results in the last column, "Average Growth Score," provide the results for all students statewide for each performance level. Thus, for example, the average growth score earned in 2003 by all students performing at Level I in 2002 was 120.5, while the average growth score earned by all students performing at Level V was 63.5. The lower the performance level in 2002, the higher the number of growth points the student was likely to earn. For the two extreme levels, the result is rather obvious; Level I students cannot earn fewer than 100 points, while Level V students cannot earn more than 100 points. As a result, the average for Level I students has to be higher than that for Level V students. But the trend holds up even for the middle levels; the average growth score for students at any level is lower than the comparable score for students at a lower level.

Table 2

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 |

It was obvious that regression was playing a role here that needed to be taken into account. However, once we started looking at the potential of using Value Tables, it became clear that an

additional possibility could—and should—be taken into account. Rather than valuing all gains and losses equally, it would be possible to value some outcomes more highly than others. For example, under NCLB, all gains below Level III are considered inconsequential, gains from below Level III to above Level III are highly valued, and gains above Level III are once again considered inconsequential (Note that some can—and do—have values that differ greatly from those of NCLB, but at least NCLB has clearly stated what its goals are). In contrast, the Alaska state legislature is considering awarding cash rewards to schools based on the changes in student performance from year to year, and the proposed legislation specifically states that schools should be rewarded, in part, for their success in moving Proficient students to Advanced, and maintaining students who already are Advanced at that level. Regardless of how one might feel about the contrasting priorities of NCLB and Alaska, it is clear that the value systems behind each piece of legislation are considerably different, and therefore the Value Tables that reflected each set of values would be just as different. Thus, the use of Value Tables not only permitted, but indeed required, that policy-makers explicitly state what educational outcomes they valued most highly, and that the Value Table that should be used in each state needed to be tailored to the values of the policy makers.

As a result, we returned to the state and reviewed our first proposed Value Table with state policy makers. During that meeting, they 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 3 (the Average Growth Score is appended to each column).

Table 3

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

| Year 1 Performance Level | Year 2 Performance Level | | | | | Average Growth 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 B overshoots the mark when regression is taken 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 4.

Table 4

Another 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 | 400 | 400 | 400 | 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 an appropriate reflection of that fact.

**Correlations with Status Scores and Reliability**

If growth scores are supposed to be a measure of the effectiveness of schools, and it is presumed that schools in higher socio-economic areas (which have higher-achieving students) provide, on average, more effective schooling than schools in lower socio-economic areas, then there should be some moderate positive correlation between the status scores of schools and their growth scores. Researchers conducting HLM studies have reported negative correlations between status and the measures they are computing, which raise questions about the validity of those results for measuring school effectiveness. The correlations between status and growth scores were -0.23 when Table A was applied, +0.61 for Table B, and +0.44 for Table C. That last result seems to be the most reasonable, lending further support to the validity of the Table C results.

We also looked at the reliability of the growth scores using Table C. We computing multiple scores for each school by drawing samples with replacement and compared the values we got for schools under pairs of draws for all schools with 20 students or more. The reliability of status scores was 0.99. We had previously looked at the stability of improvement scores (comparing the status scores attained by one cohort with the status scores obtained by the next year's cohort) under similar conditions and found the reliability of the improvement scores to be 0.87. The reliability of the growth scores was 0.94—a value considerably lower than the status scores, but also considerably higher than the improvement scores. This is an additional indication that measures of growth may be more appropriate for school-level accountability than improvement scores.

**Comparisons of Different Value Tables and of Value Tables to Other Measures of Growth**

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 5.

Table 5

Correlations among Several Measures of School Growth

| | ANCOVA | HLM Slope | Value Table A | Value Table B | Value Table C |
|---|---|---|---|---|---|
| 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 A | | | | .54 | .69 |
| Value Table B | | | | | .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 C 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. Again, if these statistics are supposed to measure teacher effectiveness, and policy makers believe the there is a moderate tendency for better teachers to be located in higher scoring schools, then these results suggest that the statistics being calculated by ANCOVA and HLM Slope are not the ones policy makers would choose if they truly understood the procedures.

**Next Steps**

Our initial analyses show relatively low correlations among school growth scores depending on the Value Table chosen. We also know that some Value Tables that appear on their surface to be appropriate turn out to be poorly correlated with other school-level statistics that should be indicating school effectiveness. This suggests that the process for establishing the Value Table to be used in a state school-level accountability system needs to be better understood than it does now. For example, we have developed a procedure similar to standard setting that allows policy-makers to articulate the values they wish to see reflected in their accountability system. We do not yet know how to create a Value Table that accurately reflects those values.

While our approach to school-level accountability is much simpler than HLM models, it uses just one year's worth of prior test scores (and just one content area from that one year). We know that researchers have shown that student-level scores are much more accurately predicted when one uses multiple years of data from several content areas. Thus, it almost certainly is true that Value Tables would not be an adequate substitute for HLM procedures if the goal were

student-level predictions. However, we also know that the prediction errors tend to average out when results are reported at the teacher and school levels, so it is unclear whether the added complexity of HLM models provides significantly better predictability at those levels. Therefore, our next step will be to collect student-level data from another state for which more sophisticated HLM predictions have been made. We will apply a series of Value Tables to those same data and compare the correlations, at the student, teacher and school level, much as we have done in Table 6 in this paper. Only when we have completed such a study will we know to what extent the added complexity of HLM models changes the measures of growth from this simple model, and of these simple models, which ones best parallel the HLM models. Such information should suggest an appropriate course of action for policy makers.

Finally, to the extent that measures of school growth generated from some Value Tables correlate highly with the results of HLM models, a careful look should be taken at the outliers. We would need to question why some schools have a high relative rank under one model and a low rank under another. The answers to those questions should shed light on which of the statistical procedures is better at ranking out schools according to the values established by the state policy-makers.