WHITE PAPER

SAS® EVAAS® for K-12 Statistical Models

THE POWER TO KNOW.

Contents

1 Introduction

For the past two decades, the team behind SAS® EVAAS® for K-12 has provided analytic services to educators and policymakers regarding the effectiveness of schooling practices. In the early years, focusing on students' growth over time, rather than their initial achievement, represented a paradigm shift in identifying effective schooling and teaching. Furthermore, by following the progress of individual students over time, the analytics represented the philosophy that all students, regardless of their initial achievement, deserve to make appropriate academic progress each year. Today, the EVAAS services include value-added reports, individual student projections, diagnostic reports, and more, all available through an interactive web application. This document provides details on the statistical models used in the EVAAS value-added and projection analyses.

It is important to keep in mind that there is not one, single EVAAS model used in all applications. There are multiple models implemented according to the objectives of the analyses, the characteristics and availability of the test data, and the policies and preferences of educators and policymakers. However, EVAAS typically uses two general types of value-added models, which are described conceptually and technically in this document.

In addition to value-added modeling, EVAAS provides projected scores for individual students on tests the students have not yet taken. These tests may include state-mandated tests (end-of-grade tests, end-of-course tests where available) as well as national tests such as college entrance exams (SAT and ACT). These projections can be used to predict a student's future success (or lack of success) and may be used to guide counseling and intervention.

Each of these analytic methodologies will be discussed in depth throughout the remainder of this document, which is organized as follows.

- Section 2 describes the data requirements for value-added and projection reporting.
- Section 3 describes the value-added modeling approaches.
- Section 4 describes the projection modeling approach.

2 Data Requirements

2.1 Tests used by EVAAS

EVAAS analyses can make use of a wide variety of assessments including, but not limited to, state criterion referenced tests, national norm referenced tests, college ready assessments, and even some locally developed state- or district-based tests, such as career and technical education or vocational tests.

Tests are examined each year to determine if they are appropriate to use in a longitudinally linked analysis. Scales must meet the three requirements described below to be used in all types of analysis done within EVAAS. Stretch and reliability are checked every year using the distribution of scale scores that are used each year.

2.1.1 Stretch

Stretch indicates whether the scaling of the test permits student growth to be measured for either very low- or very high-achieving students. A test “ceiling” or “floor” inhibits the ability to assess growth for students who would have otherwise scored higher or lower than the test allowed. There must be enough test scores at the high or low end of achievement for a measurable difference to be observed. Stretch can be determined by the percentage of students who score near the minimum or the maximum level for each assessment. As an example, if a much larger percentage of students scored at the maximum in one grade compared to the prior grade, then it may seem that these students had negative growth at the very top of the scale. However, this is likely due to the artificial ceiling of the assessment.

2.1.2 Relevance

Relevance indicates whether the test is aligned with the curriculum. Tested material will correlate with standards if the assessments are designed to assess what students are expected to know and be able to do at each grade level. Generally, this is determined by the state or district implementing the assessments.

2.1.3 Reliability

Reliability can be viewed in a few different ways for assessments. Psychometricians view reliability as the idea that a student would receive similar scores if they took the assessment multiple times. Reliability also refers to the assessment’s scales across years. Both types of reliability are important when measuring growth. The first type of reliability is important for most any use of standardized assessments. The second type of reliability is very important when a base year is used to set the expectation of growth since this approach assumes that scale scores mean the same thing in a given subject and grade across years.

3 EVAAS Value-Added Models

Conceptually, growth compares the entering achievement of a group of students to their current achievement. Value-added models measure the amount of growth a group of students is making and attributes it up to the district, school or teacher level. The value-added model compares the growth for that group to an expected amount of growth and can provide information as to whether there is statistical evidence that the group of students exceeded, met, or did not meet that expectation.

In practice, growth must be measured using an approach that is sophisticated enough to accommodate many non-trivial issues associated with student testing data. Such issues include students with missing test scores, students with differing entering achievement, and measurement error in the test. EVAAS provides two general types of value-added models, each comprised of district-, school-, and teacher-level reports.

- **Multivariate Response Model (MRM**) can be used for tests given in consecutive grades, like the math and reading tests often implemented in grades three through eight.
- **Univariate Response Model (URM)** is used when a test is given in non-consecutive grades, or it can be used for any type of testing scenario.

Both models offer the following advantages:

- The models include all of each student's testing history without imputing any test scores.
- The models can accommodate students with missing test scores.
- The models can accommodate team teaching or other shared instructional practices.
- The models can use all years of student testing data to minimize the influence of measurement error.
- The models can accommodate tests on different scales.

Each model is described in greater detail below throughout this section.

The models described in this document only include student test scores as inputs. As a result of using all available test scores and including students, even if they have missing test scores, it is not necessary to make *direct* adjustments for students' background characteristics. In short, each student serves as his or her own control and, to the extent that socioeconomic/demographic influences persist over time, these influences are already represented in the student's data. In other words, while technically feasible, adjusting for student characteristics in sophisticated modeling approaches is not necessary from a statistical perspective. However, there are other policy considerations which may make this adjustment necessary, and this is possible with either modeling approach used by EVAAS.

3.1 Multivariate Response Model (MRM)

EVAAS provides three separate analyses using the MRM approach, one each for districts, schools, and teachers. The district and school models are essentially the same. They perform well with the large numbers of students that are characteristic of districts and most schools. The teacher model uses a different approach that is more appropriate with the smaller numbers of students typically found in teachers' classrooms. All three models are statistical models known as *linear mixed models* and can be further described as *multivariate, repeated-measures models*.

The MRM is a *gain-based model*, which means that it measures growth between two points in time for a group of students. The growth expectation is met when a cohort of students from grade to grade maintains the same relative position with respect to statewide student achievement in that year for a specific subject and grade.

The key advantages of the MRM approach can be summarized as follows:

- All students with valid data are included in the analyses, even if they have missing test scores. All of each student's testing history is included without imputing any test scores.
- By including all students in the analyses, even those with a sporadic testing history, it provides the most realistic estimate of achievement available.
- It minimizes the influence of measurement error inherent in academic assessments by using multiple data points of student test history.
- It allows educators to benefit from all tests, even when tests are on differing scales.
- It accommodates teaching scenarios where more than one teacher has responsibility for a student's learning in a specific subject/grade/year.
- The model analyzes all consecutive grade subjects simultaneously to improve precision and reliability.

Despite such rigor, conceptually, the MRM model is quite simple: did a group of students maintain the same relative position with respect to statewide student achievement from one year to the next for a specific subject and grade?

3.1.1 MRM at the conceptual level

An example data set with some description of possible value-added approaches may be helpful for conceptualizing how the MRM works. Assume that ten students are given a test in two different years with the results shown in Table 1 and 2. The goal is to measure academic growth (gain) from one year to the next. Two simple approaches are to calculate the mean of the differences *or* to calculate the differences of the means. When there is no missing data, these two simple methods provide the same answer (5.80 on the left in Table 1); however, when there is missing data, each method provides a different result (9.57 vs. 3.97 on the right in Table 2). A more sophisticated model is needed to address this problem.

Table 1: Scores without missing data

Student	Previous Score	Current Score	Gain
1	51.9	74.8	22.9
2	37.9	46.5	8.6
3	55.9	61.3	5.4
4	52.7	47.0	-5.7
5	53.6	50.4	-3.2
6	23.0	35.9	12.9
7	78.6	77.8	-0.8
8	61.2	64.7	3.5
9	47.3	40.6	-6.7
10	37.8	58.9	21.1
Mean	**49.99**	**55.79**	**5.80**
	Difference	**5.80**	

Table 2: Scores with missing data

Student	Previous Score	Current Score	Gain
1	51.9		
2	37.9		
3	55.9	61.3	5.4
4	52.7	47.0	-5.7
5	53.6	50.4	-3.2
6	23.0	35.9	12.9
7		77.8	
8		64.7	
9	47.3	40.6	-6.7
10	37.8	58.9	21.1
Mean	**45.01**	**54.58**	**3.97**
	Difference	**9.57**	

The MRM uses the correlation between current and previous scores in the non-missing data to estimate a mean for the set of all previous and all current scores as if there were no missing data. It does this *without* explicitly imputing values for the missing scores. The difference between these two estimated means is an estimate of the average gain for this group of students. In this small example, the estimated difference is 5.8. Even in a small example such as this, the estimated difference is much closer to the difference with no missing data than either measure obtained by the mean of the differences (9.57) or difference of the means (3.97). This method of estimation has been shown, on average, to outperform both of the simple methods.[1] In this small example, there were only two grades and one subject. Larger data sets, such as those used in actual EVAAS analyses, provide better correlation estimates by having more student data, subjects, and grades, which in turn provide better estimates of means and gains.

This small example is meant to illustrate the need for a model that will accommodate incomplete data and provide a reliable measure of progress. It represents the conceptual idea of what is done with the school and district models. The teacher model is slightly more complex, and all models are explained in more detail below (in Section 3.1.3). The first step in the MRM is to define the scores that will be used in the model.

---

1 See, for example: Wright, S. P. (2004), "Advantages of a Multivariate Longitudinal Approach to Educational Value- Added Assessment Without Imputation," Paper presented at National Evaluation Institute, on-line at http://www.createconference.org/documents/archive/2004/Wright-NEI04.pdf.

3.1.2 Normal curve equivalents

3.1.2.1 *Why EVAAS uses normal curve equivalents in MRM*

The MRM estimates academic growth as a "gain," or the difference between two measures of achievement from one point in time to the next. For such a difference to be meaningful, the two measures of achievement (that is, the two tests whose means are being estimated) must measure academic achievement on a common scale. Some test companies supply vertically scaled tests as a way to meet this requirement. A reliable alternative when vertically scaled tests are not available is to convert scale scores to normal curve equivalents (NCEs).

NCEs are on a familiar scale because they are scaled to look like percentiles. However, NCEs have a critical advantage for measuring growth: they are on an equal-interval scale. This means that for NCEs, unlike percentile ranks, the distance between 50 and 60 is the same as the distance between 80 and 90. NCEs are constructed to be equivalent to percentile ranks at 1, 50, and 99, with the mean being 50 and the standard deviation being 21.063 by definition. Although percentile ranks are usually truncated below 1 and above 99, NCEs are allowed to range below zero and above 100 to preserve their equal-interval property and to avoid truncating the test scale. Truncating would create an artificial ceiling or floor, and this could bias the results of the value-added measure for certain types of students by forcing the gain to be close to zero or even negative.

The NCEs used in EVAAS analyses are typically based on a *reference distribution*, or the distribution of scores on a state-mandated test for all students in each year. By definition, the mean (or average) NCE score for the reference distribution is 50 for each grade and subject. "Growth" is the difference in NCEs from one year/grade to the next in the same subject. The growth standard, which represents a "normal" year's growth, is defined by a value of zero. More specifically, it maintains the same position in the reference distribution from one year/grade to the next.

It is important to reiterate that a gain of zero on the NCE scale does *not* indicate "no growth." Rather, it indicates that a group of students in a district, school, or classroom has maintained the same position in the state distribution from one grade to the next. The expectation of growth can be set differently: either by using a specific year's reference distribution to create NCEs or by using each individual year to create NCEs. For more on the Growth Expectation, see Section 3.3.

3.1.2.3 *How EVAAS models use normal curve equivalents in MRM*

There are multiple ways of creating NCEs. EVAAS uses a method that does not assume the underlying scale is normal since experience has shown that some testing scales are not normally distributed and not assuming normality will ensure an equal-interval scale. Table 3 provides an example of the way that EVAAS converts scale scores to NCEs.

The first five columns of Table 3 show an example of a tabulated distribution of test scores from a sample set of data. The tabulation shows, for each possible test score, in a particular subject, grade, and year, how many students made that score ("Frequency") and what percent ("Percent") that frequency was out of the entire student population (in Table 3 the total number of students is approximately 130,000). Also tabulated are the cumulative frequency ("Cum Freq"), which is the number of students who made that score or lower, and its associated percentage ("Cum Pct").

The next step is to convert each score to a percentile rank, listed as "Ptile Rank" on the right side of Table 3. If a particular score has a percentile rank of 48, this is interpreted to mean that 48% of students in the population had a lower score and 52% had a higher score. In practice, a non-zero percentage of students will receive each specific score. For example, 3.4% of students received a score of 425 in Table

3. The usual convention is to consider half of that 3.4% to be "below" and half "above." Adding 1.7% (half of 3.4%) to the 43.5% who scored below the score of 425 produces the percentile rank of 45.2 in Table 3.

Table 3: Converting tabulated test scores to NCE values

Score	Frequency	Cum Freq	Percent	Cum Pct	Ptile Rank	Z	NCE
418	3,996	48,246	3.1	36.9	35.4	-0.375	42.10
420	4,265	52,511	3.3	40.2	38.5	-0.291	43.86
423	4,360	56,871	3.3	43.5	41.8	-0.206	45.66
425	4,404	61,275	3.4	46.9	45.2	-0.121	47.45
428	4,543	65,818	3.5	50.4	48.6	-0.035	49.26
430	4,619	70,437	3.5	53.9	52.1	0.053	51.12
432	4,645	75,082	3.6	57.5	55.7	0.142	53.00

NCEs are obtained from the percentile ranks using the normal distribution. Using a table of the standard normal distribution (found in many textbooks) or computer software (for example, a spreadsheet), one can obtain, for any given percentile rank, the associated Z-score from a standard normal distribution. NCEs are Z-scores that have been rescaled to have a "percentile-like" scale. Specifically, NCEs are scaled so that they exactly match the percentile ranks at 1, 50, and 99. This is accomplished by multiplying each Z-score by approximately 21.063 (the standard deviation on the NCE scale) and adding 50 (the mean on the NCE scale).

An alternative to normalization (using NCEs) that is sometimes suggested is "standardization." Standardized scores are scores that have been rescaled to have a mean of zero and a standard deviation of one. Just as with NCEs, scores could be standardized each year separately or to a base year. Just as for NCEs, a gain of zero represents maintaining that same relative position in student achievement *but only for the average student*. If the distributions of the two scores making up the gain have different shapes, then for students who are not average, maintaining the same relative position in student achievement may be represented by a gain of greater or less than zero. If the two distributions are both normal, then standardization produces the same results as normalization. In contrast, with NCEs a gain of zero represents maintaining the same relative position in student achievement for every student, no matter where in the distribution their score falls. It is for this reason, along with the familiarity of the NCE scale, that EVAAS analyses use NCE scores.

3.1.3 Technical description of the linear mixed model and the MRM

The linear mixed model for district, school, and teacher value-added reporting using the MRM approach is represented by the following equation in matrix notation:

$$y = X\beta + Zv + \epsilon \qquad (1)$$

y (in the EVAAS context) is the $m \times 1$ observation vector containing test scores (usually NCEs) for all students in all academic subjects tested over all grades and years.

X is a known $m \times p$ matrix that allows the inclusion of any fixed effects.

β is an unknown $p \times 1$ vector of fixed effects to be estimated from the data.

Z is a known $m \times q$ matrix that allows for the inclusion of random effects.

v is a non-observable $q \times 1$ vector of random effects whose realized values are to be estimated from the data.

ϵ is a non-observable $m \times 1$ random vector variable representing unaccountable random variation.

Both v and ϵ have means of zero, that is, $E(v = 0)$ and $E(\epsilon = 0)$. Their joint variance is given by:

$$Var\begin{bmatrix} v \\ \epsilon \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \quad (2)$$

where R is the $m \times m$ matrix that reflects the correlation among the student scores residual to the specific model being fitted to the data, and G is the $q \times q$ variance-covariance matrix that reflects the correlation among the random effects. If (v, ϵ) are normally distributed, the joint density of (y, v) is maximized when β has value b and v has value u given by the solution to the following equations, known as Henderson's mixed model equations (Sanders et al., 1997):

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} \quad (3)$$

Let a generalized inverse of the above coefficient matrix be denoted by:

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix}^{-} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = C \quad (4)$$

If G and R are known, then some of the properties of a solution for these equations are:

1. Equation (5) below provides the best linear unbiased estimator (BLUE) of the set of estimable linear function,$K^T\beta$, of the fixed effects. The second equation (6) below represents the variance of that linear function. The standard error of the estimable linear function can be found by taking the square root of this quantity.

$$E(K^T\beta) = K^T b \quad (5)$$

$$Var(K^T b) = (K^T)C_{11}K \quad (6)$$

2. Equation (7) below provides the best linear unbiased predictor (BLUP) of v.

$$E(v|u) = u \quad (7)$$

$$Var(u - v) = C_{22} \quad (8)$$

   where u is unique regardless of the rank of the coefficient matrix.

3. The BLUP of a linear combination of random and fixed effects can be given by equation (9) below provided that $K^T\beta$ is estimable. The variance of this linear combination is given by equation (10).

$$E(K^T\beta + M^T v\,|u) = K^T b + M^T u \quad (9)$$

$$Var(K^T(b - \beta) + M^T(u - v)) = (K^T M^T)C(K^T M^T)^T \quad (10)$$

4. With G and R known, the solution for the fixed effects is equivalent to generalized least squares, and if v and ϵ are multivariate normal, then the solutions for β and v are maximum likelihood.

5. If G and R are not known, then as the estimated G and R approach the true G and R, the solution approaches the maximum likelihood solution.

6. If v and ϵ are not multivariate normal, then the solution to the mixed model equations still provides the maximum correlation between v and u.

3.1.3.1 *District and school level*

The district and school MRMs do not contain random effects; consequently, in the linear mixed model, the Zv term drops out. The X matrix is an incidence matrix (a matrix containing only zeros and ones) with a column representing each interaction of school (in the school model), subject, grade, and year of data. The fixed-effects vector β contains the mean score for each school, subject, grade, and year, with each element of β corresponding to a column of X. Note that, since MRMs are generally run with each school uniquely defined across districts, there is no need to include district in the school model.

Unlike the case of the usual linear model used for regression and analysis of variance, the elements of ϵ are *not* independent. Their interdependence is captured by the variance-covariance matrix, also known as the R matrix. Specifically, scores belonging to the same student are correlated. If the scores in y are ordered so that scores belonging to the same student are adjacent to one another, then the R matrix is block diagonal with a block, R_i, for each student. Each student's R_i is a subset of the "generic" covariance matrix R_0 that contains a row and column for each subject and grade. Covariances among subjects and grades are assumed to be the same for all years (technically, all cohorts), but otherwise the R_0 matrix is unstructured. Each student's R_i contains only those rows and columns from R_0 that match the subjects and grades for which the student has test scores. In this way, the MRM is able to use all available scores from each student.

Algebraically, the district MRM is represented as:

$$y_{ijkld} = \mu_{jkld} + \epsilon_{ijkld} \tag{11}$$

where y_{ijkld} represents the test score for the i^{th} student in the j^{th} subject in the k^{th} grade during the l^{th} year in the d^{th} district. μ_{ijkld} is the estimated mean score for this particular district, subject, grade, and year. ϵ_{ijkld} is the random deviation of the i^{th} student's score from the district mean.

The school MRM is represented as:

$$y_{ijkls} = \mu_{jkls} + \epsilon_{ijkls} \tag{12}$$

This is the same as the district analysis with the replacement of subscript d with subscript s representing the s^{th} school.

The MRM uses the data for the up to the most recent eight years each year to estimate the covariances that can be found in the matrix R_0. This estimation of covariances is done within each level of analyses and can result in slightly different values within each analysis.

Solving the mixed model equations for the district or school MRM produces a vector b that contains the estimated mean score for each school (in the school model), subject, grade, and year. To obtain a value-added measure of average student growth, a series of computations can be done using the students from a school in a particular year and all of their prior year schools. Because students may change schools from one year to the next (in particular when transitioning from elementary to middle school, for example), the estimated mean score for the prior year/grade utilizes a weighted average of schools that sent students into the school, grade, subject, and year in question. Prior year schools are not typically utilized if they send students in very small amounts (less than five) since those students likely do not represent the overall achievement of the school that they are coming from. For certain schools with very large rates of mobility, the estimated mean for the prior year/grade only includes students who existed in the current year. Mobility is taken into account within the model so that growth of students is computed using all students in each school, including those that may have moved buildings from one year to the next.

The computation for obtaining a growth measure can be thought of as a linear combination of fixed effects from the model. The best linear unbiased estimate for this linear combination is given by equation (5). The growth measures are reported along with standard errors, and these can be obtained by taking the square root of equation (6).

Furthermore, in addition to reporting the estimated mean scores and mean gains produced by these models, the value-added reporting can include (1) cumulative gains across grades (for each subject and year) and (2) multi-year average gains (up to three years for each subject and grade). In general, these are all different forms of linear combinations of the fixed effects, and their estimates and standard errors are computed in the same manner described above.

3.1.3.2 *Teacher-level*

The teacher estimates use a more conservative statistical process to lessen the likelihood of misclassifying teachers. Each teacher is assumed to be the state or district average in a specific year, subject, and grade until the weight of evidence pulls him or her above or below that average. Furthermore, the teacher model is a “layered” model, which means that:

- The model incorporates current and previous teacher effects.
- Each teacher estimate takes into account all the students’ testing data over the years.
- The model incorporates the percentage of instructional responsibility that a teacher has for each student (to accommodate scenarios such as team teaching).

Each of these elements of the statistical model for teacher value-added modeling provides a layer of protection against misclassifying each teacher estimate.

To allow for the possibility of many teachers with relatively few students per teacher, MRM enters teachers as random effects via the Z matrix in the linear mixed model. The X matrix contains a column for each subject/grade/year, and the b vector contains an estimated state or district mean score for each subject/grade/year. The Z matrix contains a column for each subject/grade/year/teacher, and the u vector contains an estimated teacher effect for each subject/grade/year/teacher. The R matrix is as described above for the district or school model. The G matrix contains teacher variance components, with a separate unique variance component for each subject/grade/year. To allow for the possibility that a teacher may be very effective in one subject and very ineffective in another, the G matrix is constrained to be a diagonal matrix. Consequently, the G matrix is a block diagonal matrix with a block for each subject/grade/year. Each block has the form $\sigma^2{}_{jkl}I$ where $\sigma^2{}_{jkl}$ is the teacher variance component for the j^{th} subject in the k^{th} grade in the l^{th} year, and I is an identity matrix.

Algebraically, the teacher model is represented as:

$$y_{ijkl} = \mu_{jkl} + \left(\sum_{k^* \leq k} \sum_{t=1}^{T_{ijk^*l^*}} w_{ijk^*l^*t} \times \tau_{ijk^*l^*t} \right) + \epsilon_{ijkl} \qquad (13)$$

y_{ijkl} is the test score for the i^{th} student in the j^{th} subject in the k^{th}grade in the l^{th} year. $\tau_{ijk^*l^*t}$ is the teacher effect of the t^{th} teacher on the i^{th} student in the j^{th} subject in grade k^* in year l^*.

The complexity of the parenthesized term containing the teacher effects is due to two factors. First, in any given subject/grade/year, a student may have more than one teacher. The inner (rightmost) summation is over all the teachers of the i^{th} student in a particular subject/grade/year. $\tau_{ijk^*l^*t}$ is the effect of the t^{th} teacher. $w_{ijk^*l^*t}$ is the fraction of the i^{th} student's instructional time claimed by the t^{th} teacher. Second, as mentioned above, this model allows teacher effects to accumulate over time. That is, how well a student does in the current subject/grade/year depends not only on the current teacher but also on the accumulated knowledge and skills acquired under previous teachers. The outer (leftmost) summation accumulates teacher effects not only for the current (subscripts k and l) but also over previous grades and years (subscripts k^* and l^*) in the same subject. Because of this accumulation of teacher effects, this type of model is often called the "layered" model.

In contrast to the model for many district and school estimates, the value-added estimates for teachers are not calculated by taking differences between estimated mean scores to obtain mean gains. Rather, this teacher model produces teacher "effects" (in the u vector of the linear mixed model). It also produces, in the fixed-effects vector b, state-level or district-level mean scores (for each year, subject and grade). Because of the way the X and Z matrices are encoded, in particular because of the "layering" in Z, teacher gains can be estimated by adding the teacher effect to the state or district mean gain. That is, the interpretation of a teacher effect in this teacher model is as a gain, expressed as a deviation from the average gain for the state in a given year, subject, and grade.

Table 4 illustrates how the Z matrix is encoded for three students who have three different scenarios of teachers during grades three, four, and five in two subjects, math (M) and reading (R). Teachers are identified by the letters A–F.

Tommy's teachers represent the conventional scenario: Tommy is taught by a single teacher in both subjects each year (teachers A, C, and E in grades three, four, and five, respectively). Notice that in Tommy's Z matrix rows for grade four, there are ones (representing the presence of a teacher effect) not only for fourth grade teacher C but also for third grade teacher A. This is how the "layering" is encoded. Similarly, in the grade five rows, there are ones for grade five teacher E, grade four teacher C, and grade three teacher A.

Susan is taught by two different teachers in grade three, teacher A for math and, teacher B for reading. In grade four, Susan had teacher C for reading. For some reason in grade four, no teacher claimed Susan for math even though Susan had a grade four math test score. This score can still be included in the analysis by entering zeroes into the Susan's Z matrix rows for grade four math. In grade five, on the other hand, Susan had no test score in reading. This row is completely omitted from the Z matrix. There will always be a Z matrix row corresponding to each test score in the y vector. Since Susan has no entry in y for grade five reading, there can be no corresponding row in Z.

Eric's scenario illustrates team teaching. In grade three reading, Eric received an equal amount of instruction from both teachers A and B. The entries in the Z matrix indicate each teacher's contribution, 0.5 for each teacher. In grade five math, however, while Eric was taught by both teachers E and F, they did not make an equal contribution. Teacher E claimed 80% responsibility, and teacher F claimed 20%.

Because teacher effects are treated as random effects in this approach, their estimates are obtained by shrinkage estimation, technically known as best linear unbiased prediction or as empirical Bayesian estimation. This means that *a priori* a teacher is considered to be "average" (with a teacher effect of zero) until there is sufficient student data to indicate otherwise. This method of estimation protects against false positives (teachers incorrectly evaluated as effective) and false negatives (teachers incorrectly evaluated as ineffective), particularly in the case of teachers with few students.

From the computational perspective, the teacher gain can be defined as a linear combination of both fixed effects and random effects and is estimated by the model using equation (9). The variance and standard error can be found using equation (10). As described in the district/school section, the teacher value-added reporting can include (1) cumulative gains across grades (for each subject and year) and (2) multi-year average gains (up to three years for each subject and grade). In general, these are all different forms of linear combinations of the fixed effects, and their estimates and standard errors are computed in the same manner described above.

Table 4: Encoding the Z matrix

			Teachers											
			Third Grade				Fourth Grade				Fifth Grade			
			A		B		C		D		E		F	
Student	**Grade**	**Subjects**	**M**	**R**	**M**	**R**	**M**	**R**	**M**	**R**	**M**	**R**	**M**	**R**
Tommy	**3**	**M**	1	0	0	0	0	0	0	0	0	0	0	0
		R	0	1	0	0	0	0	0	0	0	0	0	0
	4	**M**	1	0	0	0	1	0	0	0	0	0	0	0
		R	0	1	0	0	0	1	0	0	0	0	0	0
	5	**M**	1	0	0	0	1	0	0	0	1	0	0	0
		R	0	1	0	0	0	1	0	0	0	1	0	0
Susan	**3**	**M**	1	0	0	0	0	0	0	0	0	0	0	0
		R	0	0	0	1	0	0	0	0	0	0	0	0
	4	**M**	1	0	0	0	0	0	0	0	0	0	0	0
		R	0	0	0	1	0	1	0	0	0	0	0	0
	5	**M**	1	0	0	0	0	0	0	0	0	0	1	0
Eric	**3**	**M**	1	0	0	0	0	0	0	0	0	0	0	0
		R	0	0.5	0	0.5	0	0	0	0	0	0	0	0
	4	**M**	1	0	0	0	0	0	1	0	0	0	0	0
		R	0	0.5	0	0.5	0	0	0	1	0	0	0	0
	5	**M**	1	0	0	0	0	0	1	0	0.8	0	0.2	0
		R	0	0.5	0	0.5	0	0	0	1	0	1	0	0

3.2 Univariate Response Model (URM)

Tests that are not given for consecutive years require a different modeling approach from the MRM, and this modeling approach is called the univariate response model (URM). This approach can be used for tests given in consecutive grades too. The statistical model can also be classified as a linear mixed model and can be further described as an analysis of covariance (ANCOVA) model. The URM is a regression-based model, which measures the difference between students' predicted scores for a particular subject/year with their observed scores. The growth expectation is met when students with a district/school/teacher made the same amount of progress as students in the average district/school/teacher with the state for that same year/subject/grade. If not all teachers were administering a particular test in the state, then it would be compared to the average of those teachers with students taking that assessment.

The key advantages of the URM approach can be summarized as follows:

- It does not require students to have all predictors or the same set of predictors, so long as a student has at least three prior test scores in any subject/grade.
- It minimizes the influence of measurement error by using all prior data for an individual student. Analyzing all subjects simultaneously increases the precision of the estimates.
- It allows educators to benefit from all tests, even when tests are on differing scales.
- It accommodates teaching scenarios where more than one teacher has responsibility for a student's learning in a specific subject/grade/year.

3.2.1 URM at the conceptual level

The URM is run for each individual year, subject and grade (if relevant). Consider all students who took grade eight science in a given year. Those students are connected to all of their prior testing history (all grades, subjects, and years), and the relationship between the observed grade eight science scores with all prior test scores is examined. It is important to note that some prior test scores are going to have a greater relationship to the score in question than others. For instance, it is likely that prior science tests will have a greater relationship with science than prior reading scores. However, the other scores do still have a statistical relationship.

Once that relationship has been defined, a predicted score can be calculated for each individual student based on his or her own prior testing history. Of course, some prior scores will have more influence than others in predicting certain scores based on the observed relationship across the state or testing pool in a given year. With each predicted score based on a student's prior testing history, this information can be aggregated to the district, school, or teacher level. The predicted score can be thought of as the entering achievement of a student.

The measure of growth is a function of the difference between the observed (most recent) scaled scores and predicted scaled scores of students associated with each district, school, or teacher. If students at a school typically outperform their individual growth expectation, then that school will likely have a larger value-added measure. Zero is defined as the average district, school, or teacher in terms of the average progress, so that if every student obtained their predicted score, a district, school, or teacher would likely receive a value-added measure close to zero. A negative or zero value does not mean "zero growth" since this is all relative to what was observed in the state (or pool) that year.

3.2.2 Technical description of the district, school and teacher models

The URM has similar models for district and school and a slightly different model for teachers that allows multiple teachers to share instructional responsibility. The statistical details for the teacher model are outlined below.

In this model, the score to be predicted serves as the response variable (y), the dependent variable), the covariates (x's, predictor variables, explanatory variables, independent variables) are scores on tests the student has already taken, and the categorical variable (class variable, factor) are the teacher(s) from whom the student received instruction in the subject/grade/year of the response variable (y). For the district and school models, the categorical variable would be the district or school. Algebraically, the model can be represented as follows for the i^{th} student when there is no team teaching.

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i \tag{14}$$

In the case of team teaching, the single α_j is replaced by multiple α's, each multiplied by an appropriate weight, similar to the way this is handled in the teacher MRM in equation (13). The μ terms are means for the response and the predictor variables. α_j is the teacher effect for the j^{th} teacher, the teacher who claimed responsibility for the i^{th} student. The β terms are regression coefficients. Predictions to the response variable are made by using this equation with estimates for the unknown parameters (μ's, β's, sometimes α_j). The parameter estimates (denoted with "hats," e.g., $\hat{\mu}$, $\hat{\beta}$) are obtained using all of the students that have an observed value for the specific response and have three predictor scores. The resulting prediction equation for the i^{th} student is as follows:

$$\hat{y}_i = \hat{\mu}_y + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots \tag{15}$$

Two difficulties must be addressed in order to implement the prediction model. First, not all students will have the same set of predictor variables due to missing test scores. Second, the estimated parameters are pooled-within-teacher estimates. The strategy for dealing with missing predictors is to estimate the joint covariance matrix (call it C) of the response and the predictors. Let C be partitioned into response (y) and predictor (x) partitions, that is:

$$C = \begin{bmatrix} c_{yy} & c_{yx} \\ c_{xy} & C_{xx} \end{bmatrix} \tag{16}$$

Note that C in equation (16) is not the same as C in equation (4). This matrix is estimated using an EM algorithm for estimating covariance matrices in the presence of missing data such as the one provided by the MI procedure in SAS/STAT®, but modified to accommodate the nesting of students within teachers. Only students who had a test score for the response variable in the most recent year and who had at least three predictor variables are included in the estimation. Given such a matrix, the vector of estimated regression coefficients for the projection equation (15) can be obtained as:

$$\hat{\beta} = C_{xx}^{-1} c_{xy} \tag{17}$$

This allows one to use whichever predictors a particular student has to get that student's projected y-value ($\hat{y}_i$). Specifically, the C_{xx} matrix used to obtain the regression coefficients *for a particular student* is that subset of the overall C matrix that corresponds to the set of predictors for which this student has scores.

The prediction equation also requires estimated mean scores for the response and for each predictor (the $\hat{\mu}$ terms in the prediction equation). These are not simply the grand mean scores. It can be shown that in an ANCOVA, if one imposes the restriction that the estimated teacher effects should sum to zero (that is, the teacher effect for the "average teacher" is zero), then the appropriate means are the means of the teacher-level means. The teacher-level means are obtained from the EM algorithm, mentioned above, which takes into account missing data. The overall means ($\hat{\mu}$ terms) are then obtained as the simple average of the teacher-level means.

Once the parameter estimates for the prediction equation have been obtained, predictions can be made for any student with any set of predictor values, so long as that student has a minimum of three prior test scores.

$$\hat{y}_i = \hat{\mu}_y + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots \tag{18}$$

The $\hat{y}_i$ term is nothing more than a composite of all the student's past scores. It is a one-number summary of the student's level of achievement prior to the current year. The different prior test scores making up this composite are given different weights (by the regression coefficients, the $\hat{\beta}$'s) in order to maximize its correlation with the response variable. Thus a different composite would be used when the response variable is math than when it is reading, for example. Note that the $\hat{\alpha}_j$ term is not included in the equation. Again, this is because $\hat{y}_i$ represents prior achievement, before the effect of the current district, school, or teacher. To avoid bias due to measurement error in the predictors, composites are obtained only for students who have at least three prior test scores.

The second step in the URM is to estimate the teacher effects (α_j) using the following ANCOVA model:

$$y_i = \gamma_0 + \gamma_1 \hat{y}_i + \alpha_j + \epsilon_i \tag{19}$$

In the URM model, the effects (α_j) are considered to be random effects. Consequently the $\hat{\alpha}_j$'s are obtained by shrinkage estimation (empirical Bayes). The regression coefficients for the ANCOVA model are given by the γ's.

3.3 Growth Expectations in Value-added Analyses

Conceptually, growth compares the entering achievement of students to the current achievement. Value-added models measure the amount of growth a group of students is making and attributes it to the district, school, or teacher level. The value-added measure compares that growth of a group of students to an expected amount of growth, and it is very important to define that expectation.

Mathematically, the "expected" growth is typically set at zero, such that *positive* gains or effects are evidence that students made *more* than the expected progress and *negative* gains or effects are evidence that students made *less* than the expected progress.

However, the definition of "expected growth" varies by model, and the precise definition depends on the selected model and state or district preference, and this section provides more details on the options and selections for defining expected growth. Generally, expected growth can be defined as either a "base year" or an "intra-year" approach. Base year refers to a growth expectation that is based on a particular year, say 2006, and any growth in the current year will be compared to the distribution of student scores in the base year. Intra-year refers to a growth expectation that is always based on the current year (2012 for 2012 growth estimates, 2013 for 2013 growth estimates, and so on).

3.3.1 Base year approach

3.3.1.1 *Description*

The base year growth expectation is based on a cohort of students moving from grade to grade and maintaining the same relative position with respect to the statewide student achievement in the base year for a specific subject and grade.

As a simplified example, if students' achievement was at the 50th NCE in 2006 grade four math, based on the 2006 grade four math scale score distribution, and at the 52nd NCE in 2007 grade five, based on the 2006 grade five math scale score distribution, then their estimated mean gain is 2 NCEs.

The key feature is that, in theory, all educational entities could exceed or fall short of the growth expectation (or standard) in a particular subject/grade/year, and the distribution of entities that are considered above or below could change over time.

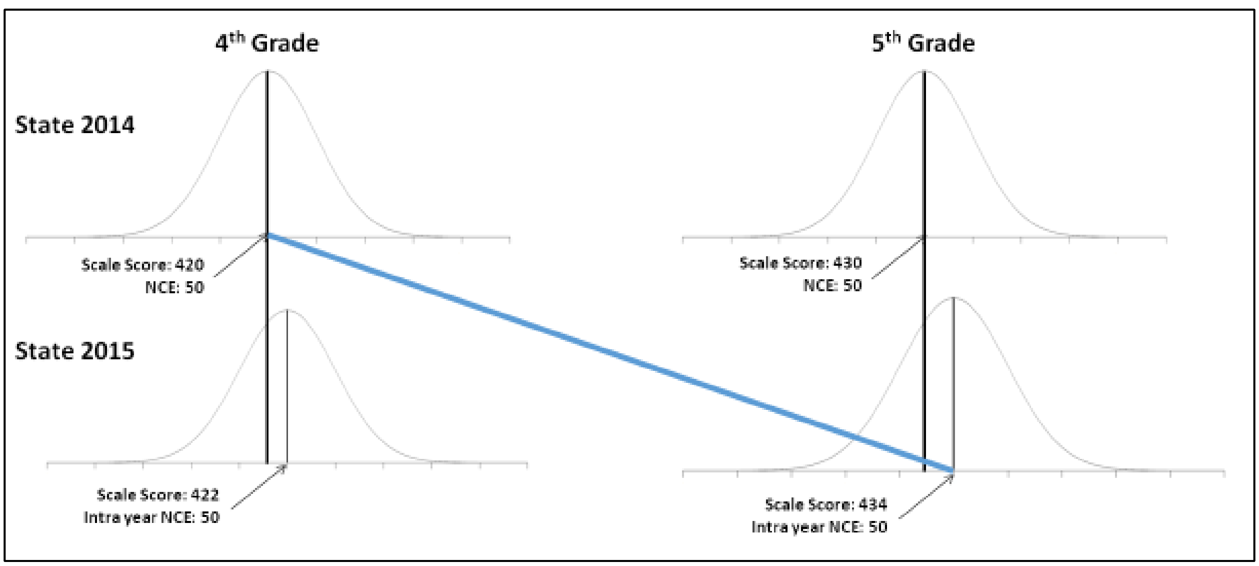
Following the implementation of any new assessments and changes in academic standards, the base year should be reset to an intra-year approach in order to accommodate the differences between the old and new testing regimes and minimize any impact on the value-added reporting. To be more specific, use of the intra-year approach is required if there is no mapping from the old assessment's scale to the new assessment's scale. However, even if that mapping does exist, the intra-year approach should be used to prevent any unusual swings in value-added measures. If a base year approach is desired after the transition, it is recommended that the new base year be selected after the second year of the new assessment, at a minimum, and only be reset when the smooth transition can be verified.

3.3.1.2 *Illustrated example*

The graphic below (Graph 1) provides a *simplified* example of how growth is calculated with a base year approach when the state achievement increases. The graphic below has four graphs, each of which plot the NCE distribution of scale scores for a given year and grade. In this example, the base year is 2014, and the graphic shows how the gain is calculated for a group of 2014 grade four students as they become 2015 grade five students. In 2014, our grade four students score, on average, 420 scale score points on the test, which corresponds to the 50th NCE (similar to the 50th percentile). In 2015, the students score, on average, 434 scale score points on the test, which corresponds to a 52nd NCE *based on the 2014 grade five distribution of scores*. The 2015 grade five distribution of scale scores was higher than the 2014 grade five distribution of scale scores, which is why the lower right-hand graph is shifted slightly to the right. The blue line shows what is required for students to make expected growth, which would maintain their position at the 50th NCE in 2014 grade four as they become 2015 grade five students. The growth measure for these students is 2015 NCE – 2014 NCE, which would be 52 – 50 = 2. Similarly, if a group of students started out at the 35th NCE in 2014 grade four and then moved their position to the 37th NCE in 2015 grade five, they would have a gain of two NCEs as well.

Please note that the actual gain calculations are much more robust than what is presented here; as described in the previous section, the models can address students with missing data, team teaching and all available testing history. This illustration simply provides the basic concept.

Graph 1: Illustrated example of base year approach



3.3.2 Intra-year approach

3.3.2.1 *Description*

This approach will be used in the MRM reporting during the transition to new assessments and the concept is always used in the URM reporting. The actual definitions in each model are slightly different, but the concept can be considered as the average amount of progress seen across the state in a statewide implementation.

Using the URM model the definition of the expectation is that students with a district, school, or teacher made the same amount of progress as students with the average district, school, or teacher in the state for that same year/subject/grade. If not all students are taking an assessment in the state, then it may be a subset.

Using the MRM model, the definition of this type of expectation of growth is that students maintained the same relative position with respect to the statewide student achievement from one year to the next in the same subject area. As an example, if students' achievement was at the 50th NCE in 2014 grade four math, based on the 2014 grade four math statewide distribution of student achievement, and their achievement is at the 50th NCE in 2015 grade five math, based on the 2015 grade five math statewide distribution of student achievement, then their estimated gain is 0.0 NCEs.

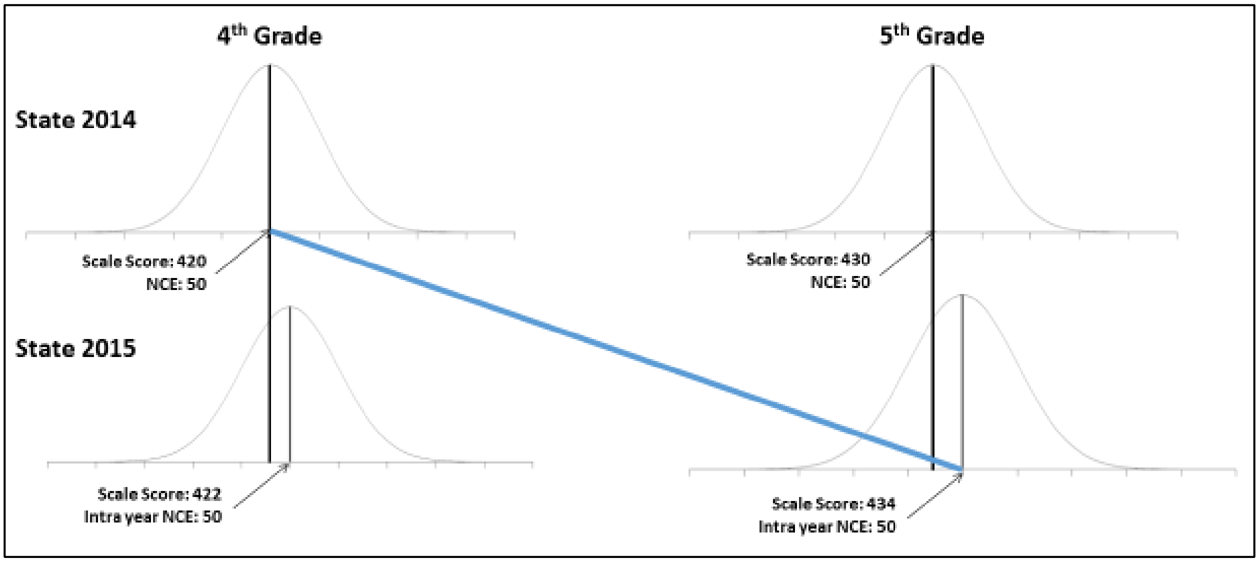
With this approach, the value-added measures tend to be centered on the growth expectation every year, with approximately half of the district/school/teacher estimates above zero and approximately half of the district/school/teacher estimates below zero. This does not mean half would be in the positive and negative categories since many value-added measures are indistinguishable from the expectation when considering the statistical certainty around that measure.

3.3.2.2 *Illustrated example*

The graphic below (Graph 2) provides a *simplified* example of how growth is calculated with an intra-year approach when the state or pool achievement increases using the MRM methodology. The graphic below has four graphs, each of which plot the NCE distribution of scale scores for a given year and grade. In this example, the first year is 2014, and the graphic shows how the gain is calculated for a group of 2014 grade four students as they become 2015 grade five students. In 2014, our grade four students score, on average, 420 scale score points on the test, which corresponds to the 50th NCE (similar to the 50th percentile). In 2015, the students score, on average, 434 scale score points on the test, which corresponds to a 50th NCE *based on the 2015 grade five distribution of scores*. The 2015 grade five distribution of scale scores was higher than the 2014 grade five distribution of scale scores, which is why the lower right-hand graph is shifted slightly to the right. The blue line shows what is required for students to make expected growth, which would maintain their position at the 50th NCE in 2014 grade four as they become 2015 grade five students. The growth measure for these students is 2015 NCE – 2014 NCE, which would be 50 – 50 = 0. Similarly, if a group of students started at the 35th NCE, the expectation is that they would maintain that 35th NCE.

Please note that the actual gain calculations are much more robust than what is presented here. As described in the previous section, the models can address students with missing data, team teaching, and all available testing history.

Graph 2: Illustrated example of intra-year approach



3.3.3 Defining the expectation of growth during an assessment change

During the change of assessments, the scales from one year to the next will be completely different from one another. This does not present any particular challenges with the URM methodology because all predictors in this approach are already on different scales from the response variable, so the transition is no different from a scaling perspective. Of course, there will be a need for the predictors to be adequately related to the response variable of the new assessment, but that typically is not an issue. With the MRM methodology, the scales from one year to the next can be completely different from one another with the intra-year approach. This method converts any scale to a relative position and can be used through an assessment change.

3.4 Using Standard Errors to Create Levels of Certainty and Define Effectiveness

In all value-added reporting, EVAAS includes the value-added estimate and its associated standard error. This section provides more information regarding standard error and how it is used to define effectiveness.

3.4.1 Using standard errors derived from the models

As described in the modeling approaches section, each model provides an estimate of growth for a district, school, or teacher in a particular subject/grade/year as well as that estimate's standard error. The standard error is a measure of the quantity and quality of student level data included in the estimate, such as the number of students and the occurrence of missing data for those students. Because measurement error is inherent in any growth or value-added model, *the standard error is a critical part of the reporting*. Taken together, the estimate and standard error provide the educators and policymakers with critical information regarding the certainty that students in a district, school, or classroom are making decidedly more or less than the expected progress. Taking the standard error into account is particularly important for reducing the risk of misclassification (for example, identifying a teacher as ineffective when he or she is truly effective) for high-stakes usage of value-added reporting.

Furthermore, because the MRM and URM models utilize robust statistical approaches as well as maximize the use of students' testing history, they can provide value-added estimates for relatively small numbers of students. This allows more teachers, schools, and districts to receive their own value-added estimates, which is particularly useful to rural communities or small schools.

The standard error also takes into account that, even among teachers with the same number of students, the teachers may have students with very different amounts of prior testing history. Due to this variation, the standard errors in a given subject/grade/year could vary significantly among teachers, depending on the available data that is associated with their students, and it is another important protection for districts, schools, and teachers to incorporate standard errors in the value-added reporting.

3.4.2 Defining effectiveness in terms of standard errors

Each value-added estimate has an associated standard error, which is a measure of uncertainty that depends on the quantity and quality of student data associated with that value-added estimate.

The standard error can help indicate whether a value-added estimate is significantly different from the growth standard. This growth standard is defined in different ways, but it is typically represented as zero on the growth scale and considered to be the *expected growth*. An index value is created that takes the growth measure and divides it by the standard error to create a t-value that can be used to discuss significance. Since the expectation of growth is zero, this measures the *certainty* about the difference of a growth measure to zero.

Many districts and schools choose to categorize the index values based on ranges to assist with their interpretation. A typical choice includes ranges of below -2, between -2 and -1, between -1 and +1, between +1 and +2, and above +2. The distribution of these categories can vary by year/subject/grade. There are many reasons this is possible, but overall, it can be shown that there are more measurable differences in some subjects and grades compared to others.

4 EVAAS Projection Model

In addition to providing value-added modeling, EVAAS provides a variety of additional services including projected scores for individual students on tests the students have not yet taken. These tests may include state-mandated tests (end-of-grade tests and end-of-course tests where available) as well as national tests such as college entrance exams (SAT and ACT). These projections can be used to predict a student's future success (or lack of success) and so may be used to guide counseling and intervention to increase students' likelihood of future success.

The statistical model that is used as the basis for the projections is, in traditional terminology, an analysis of covariance (ANCOVA) model. This model is the same statistical model used in the URM methodology applied at the school level described in Section 3.2.2. In this model, the score to be projected serves as the response variable (y), the covariates (x's) are scores on tests the student has already taken, and the categorical variable is the school at which the student received instruction in the subject/grade/year of the response variable (y). Algebraically, the model can be represented as follows for the i^{th} student.

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i \tag{20}$$

The μ terms are means for the response and the predictor variables. α_j is the school effect for the j^{th} school, the school attended by the i^{th} student. The β terms are regression coefficients. Projections to the future are made by using this equation with estimates for the unknown parameters (μ's, β's, sometimes α_j). The parameter estimates (denoted with "hats," e.g., $\hat{\mu}$, $\hat{\beta}$) are obtained using the most current data for which response values are available. The resulting projection equation for the i^{th} student is:

$$\hat{y}_i = \hat{\mu}_y \pm \hat{\alpha}_j + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots + \epsilon_i \tag{21}$$

The reason for the '±' before the $\hat{\alpha}_j$ term is that, since the projection is to a future time, the school that the student will attend is unknown, so this term is usually omitted from the projections. This is equivalent to setting $\hat{\alpha}_j$ to zero, that is, to assuming the student encounters the "average schooling experience" in the future. In some instances, a state or district may prefer to provide a list of feeder patterns from which it is possible to determine the most likely school that a student will attend at some projected future date. In this case, the $\hat{\alpha}_j$ term can be included in the projection.

Two difficulties must be addressed in order to implement the projections. First, not all students will have the same set of predictor variables due to missing test scores. Second, because of the school effect in the model, the regression coefficients must be pooled-within-school regression coefficients. The strategy for dealing with these difficulties is exactly the same as described in Section 3.2.2 using equations (16) and (17) and will not be repeated here.

Once the parameter estimates for the projection equation have been obtained, projections can be made for any student with any set of predictor values. However, to protect against bias due to measurement error in the predictors, projections are made only for students who have at least three available predictor scores. In addition to the projected score itself, the standard error of the projection is calculated ($SE(\hat{y}_i)$). Given a projected score and its standard error, it is possible to calculate the probability that a student will reach some specified benchmark of interest (b). Examples are the probability of scoring at the proficient (or advanced) level on a future end-of-grade test, or the probability of scoring sufficiently well on a college entrance exam to gain admittance into a desired program. The probability is calculated as the area above the benchmark cutoff score using a normal distribution with its mean equal to the projected score and its standard deviation equal to the standard error of the projected score as described below. Φ represents the standard normal cumulative distribution function.

$$Prob(\hat{y}_i \geq b) = \Phi\left(\frac{\hat{y}_i - b}{SE(\hat{y}_i)}\right) \tag{22}$$

To contact your local SAS office, please visit: sas.com/offices