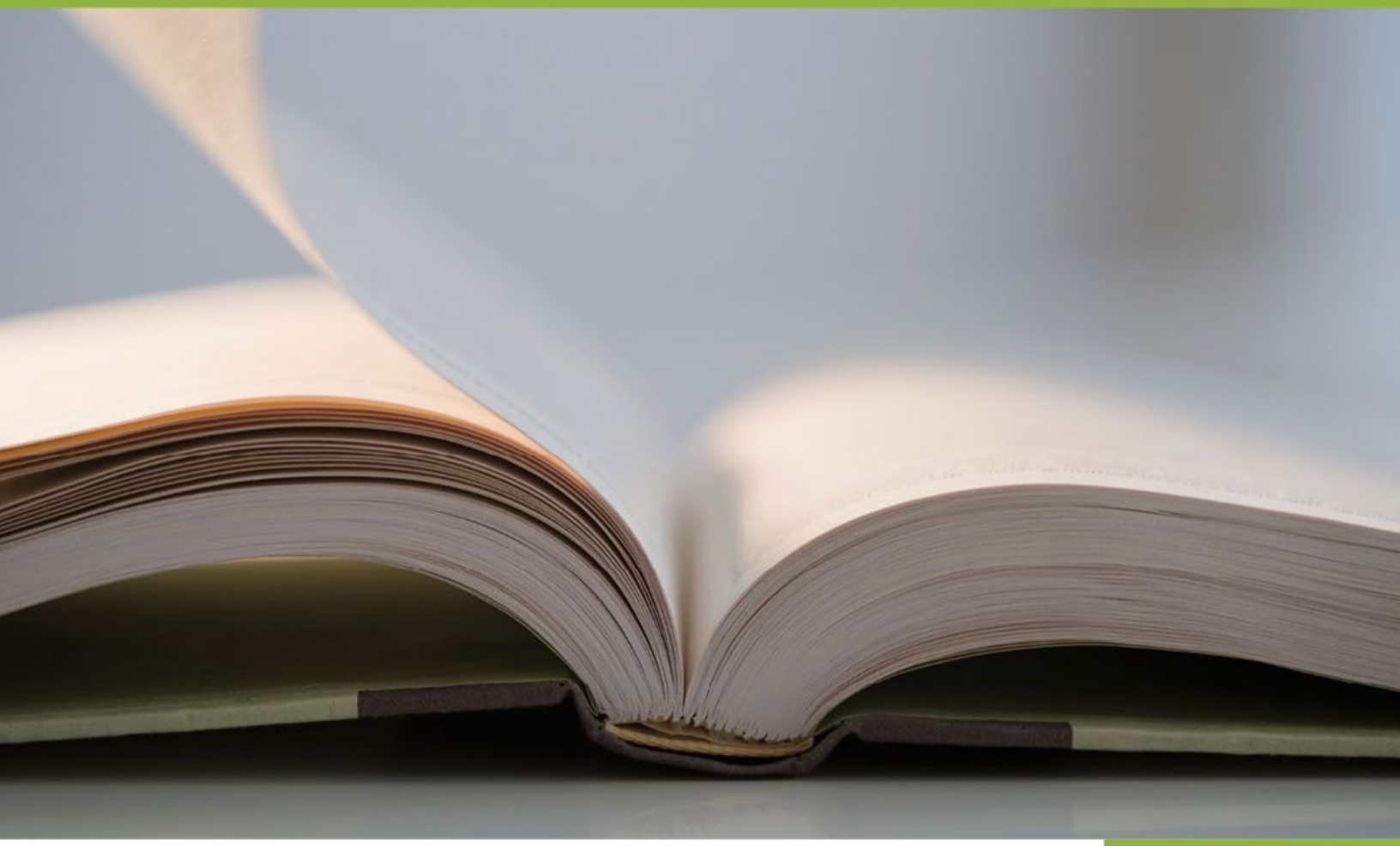
SAS® EVAAS®

Statistical Models and Business Rules of TVAAS Analyses



THE POWER TO KNOW.

Contents

1 Introduction

1.1 Value-added reporting in Tennessee

Twenty years ago, the State of Tennessee led the nation in providing measures of student progress to individual districts, schools and teachers. Known as the Tennessee Value-Added Assessment System (TVAAS), this reporting focused on the *progress* of students over time rather than their *achievement level*. TVAAS represented a paradigm shift for educators and policymakers, and in identifying the more effective practices and less effective practices, educators receive personalized feedback, which they can then leverage to improve the academic experiences of their students.

TVAAS value-added reporting began with district reporting in 1993 and expanded to school reporting in 1994 and teacher reporting in 1996.

The term "value-added" refers to a statistical analysis used to measure the amount of academic progress students make from year to year with a district, school, or teacher. Conceptually and as a simple explanation, a value-added measure is calculated in the following manner:

- Growth = current achievement/current results compared to all prior achievement/prior results with achievement being measured by a quality assessment such as the Tennessee statewide tests.

While the concept of growth is easy to understand, the implementation of a statistical model of growth is more complex. There are many decisions related to the available modeling, local policies and preferences, and business rules. Key considerations in the decision-making process include:

- What data are available?
- Given available data, what types of models are possible?
- What is the growth expectation?
- How is effectiveness defined in terms of a measure of certainty?
- What are the business rules and policy decisions that impact the way the data are processed?

The purpose of this document is to describe the value-added modeling *based on the statistical approaches, policies, and practices selected by the Tennessee Department of Education (TDOE) and currently implemented by SAS EVAAS*. This document describes the input data, modeling, and business rules for the district, school, and teacher value-added reporting in Tennessee.

1.2 What's new for 2017-18 reporting

1.2.1 Return to traditional modeling approach

During the 2015–16 school year, the State of Tennessee suspended testing in grades 3–8 for mathematics, English language arts, science, and social studies; scale scores were not available for these assessments. The 2016-17 analyses made several modeling adjustments to accommodate the missing year of data. Notably, a cumulative gain was reported in the MRM reporting, and EXPLORE data was used as a predictor for URM reporting. The 2017-18 analyses return to the traditional modeling approaches where a single-year gain will be provided for 2017-18 in the MRM reporting; EXPLORE will not be used as a predictor for URM reporting.

1.2.2 Subjects/grades that do not have reporting available

The 2017-18 reporting for districts, schools, and teachers will not include grade 2, grade 3 in science and social studies, or grade 4 in science and social studies. Grade 2 reporting is not available because there is no longer statewide testing in kindergarten and grade 1. Reporting in science and social studies for grades 3 and 4 is not available because this year's assessments have been shortened and cannot be used for value-added analysis.

1.2.3 Change in modeling for social studies

For the 2017-18 reporting year, social studies value-added measures use the predictive model (URM). The 2016-17 social studies field test validated the new social studies assessment to support full implementation during the 2017-18 school year but did not produce a full set of results that can be used as prior scores. Because 2017-18 was the first year of full implementation of the new social studies assessment, there is not a gain to measure growth from grade to grade in that subject, such as grade 6 to 7. However, the predictive model does not require a gain to measure growth and has long been used in Tennessee for assessments that are not given in consecutive grades, like the high school exams. More details about this model are available in Section 3.2.

1.2.4 Teacher evaluation composites

For the 2017-18 reporting year, there are up to three evaluation composites available for each teacher. The first is a **single-year evaluation composite** comprised solely of value-added measures from the current year reporting (i.e., 2017-18). The second is a **multi-year evaluation composite** that includes up to three years' reporting (i.e., 2015-16, 2016-17, and 2017-18) together at 35%, where 2015-16 and 2016-17 is weighted at 25% and 2017-18 at 10%. This composite is available for any teachers with 2017-18 TVAAS data and 2015-16 and/or 2016-17 TVAAS data. The third option is a **multi-year evaluation composite without 2015-16 data**, and this composite excludes any 2015-16 data and weighs 2016-17 data and 2017-18 data at 10% each. This adjusted multi-year option is available only for teachers who received 2015-16 teacher TVAAS reports.

Section 6 on page 36 provides more technical details about how these evaluation composites will be calculated.

2 Input data used in TVAAS

This Section provides details about the input data used in the Tennessee value-added model, such as the requirements for verifying appropriateness in value-added analysis as well as the student, teacher, and school information provided in the assessment files.

2.1 Determining suitability of assessments

2.1.1 Current assessments

To be used appropriately in any value-added analyses, the scales of these tests must meet three criteria. (Additional details about each of these requirements are provided in Section 8, *Data quality and pre-analytic data processing*, on page 48.)

- **There is sufficient stretch in the scales** to ensure that progress can be measured for both low-achieving students as well as high-achieving students. A floor or ceiling in the scales could disadvantage educators serving either low-achieving or high-achieving students.
- **The test is designed to assess academic standards** so that it is possible to measure progress with the assessment in that subject/grade/year. More information can be found at the following link: https://www.tn.gov/education/instruction/academic-standards.html
- **The scales are sufficiently reliable from one year to the next**. This criterion typically is met when there are a sufficient number of items per subject/grade/year, and this will be monitored each subsequent year that the test is given.

These criteria are monitored by EVAAS and psychometricians at TDOE.

The current value-added implementation in Tennessee includes many assessments measuring Tennessee's standards (TCAP Achievement, End-of-Course and Grade 2 Assessments) as well as college and career readiness assessments.

2.1.2 Transitioning to new assessments

In 2015-16, Tennessee implemented new End-of-Course (EOC) assessments in math and English language arts. Redesigned assessments in math, English language arts, and science were also implemented in grades 3-8 during the 2016-17 school year and in social studies in grades 3-8 during the 2017-18 school year. Changes in testing regimes occur at regular intervals within any state, and these changes need not disrupt the continuity and use of value-added reporting by educators and policymakers. Based on twenty years of experience with providing value-added and growth reporting to Tennessee educators, EVAAS has developed several ways to accommodate changes in testing regimes.

Prior to any value-added analyses with new tests, EVAAS verifies that the test's scaling properties are suitable for such reporting. In addition to the criteria listed above, EVAAS verifies that the new test is related to the old test to ensure that the comparison from one year to the next is statistically reliable. Perfect correlation is not required, but there should be a strong relationship between the new test and old test. For example, a new Algebra I exam should be correlated to previous math scores in grades 7 and 8 and to a lesser extent other grades and subjects such as English language arts and science. Once suitability of any new assessment has been confirmed, it is possible to use both the historical testing data and the new testing data to avoid any breaks or delays in value-added reporting.

2.2 Assessment data used in Tennessee

The state tests are administered in the spring semester except for the End-of-Course (EOC) assessments, which are given in the fall and spring semesters.

2.2.1 Tests administered in Tennessee

EVAAS receives the following tests:

- TCAP mathematics, English language arts, science, and social studies in grades 3–8.
- Grade 2 Assessments in English language arts - literature, English language arts - informational, and math in grade 2.
- EOC assessments in Algebra I, Algebra II, English I, English II, English III, Biology, Chemistry, Geometry, Integrated Math I, Integrated Math II, Integrated Math III, and U.S. History.
- ACT assessments in English, math, reading, and science.
- Advanced Placement (AP) assessments
- TCAP Alt
- Multi-State Alternate Assessment (MSAA)

2.2.2 Student identification information

The following information is received by EVAAS from TDOE:

- Student last name
- Student first name
- Student middle initial
- Student date of birth
- Student state ID number (Unique Student ID (USID))

2.2.3 Assessment information provided

EVAAS obtains all assessment information from the files provided by TDOE. These files provide the following information:

- Scale score
- Performance level
- Test taken
- Tested grade
- Tested semester
- District number
- School number
- Membership
  - School (BEEN enrolled in SCHOOL)
  - District (NOT enrolled in school but enrolled in DISTRICT)
  - State (NOT enrolled in district but enrolled in a Tennessee PUBLIC DISTRICT)
  - Not in TN (NOT enrolled in a Tennessee public district)
- Testing Status
  - Nullified
  - Medically Exempt
  - Did Not Attempt
  - Absent

- Test Form/Version/Modified Test Format
  - Large Print
  - Braille
  - ELSA
- Attendance
  - Traditional: 150 days or more
  - Traditional: 75 to 149 days
  - Traditional: 74 days or fewer
  - Block: 75 days or more
  - Block: 38 to 74 days
  - Block: 37 days or fewer

2.3 Student information

Student information is used in creating the web application to assist educators analyze the data to inform practice and assist all students with academic progress. EVAAS receives this information in the form of various socioeconomic, demographic, and programmatic identifiers provided by TDOE. Currently, these categories are as follows:

- Gifted (Not Special Ed) (Y,N)
- Gender (M,F)
- Migrant Status (Y,N)
- English Learner (Y,N)
- Title 1
  - School-wide Programs (SWP)
  - Targeted Assisted Schools (TAS)
- 504 Service Plan (Y,N)
- Economically Disadvantaged (Code AB Flag)
  - Code A: Eligible for free/reduced price lunch
  - Code B: not Eligible for free/reduced price lunch
- Special Education
  - (No) No code
  - (Yes) Less than 4 hours per week
  - (Yes) 4 through 22 hours per week
  - (Yes) More than 22 hours per week
- Functionally Delayed (Not Special Ed) (Y,N)
- Career Technical Student (High School tests only) (Y,N)
- Race
  - American Indian or Alaska Native
  - Asian
  - Black or African American
  - Hispanic
  - Multiple Races
  - Native Hawaiian/Other Pacific Islander
  - White

2.4 Teacher information

A high level of reliability and accuracy is critical for using value-added scores for both improvement purposes and high stakes decision-making. Before teacher value-added scores are calculated, teachers in Tennessee are given the opportunity to complete roster verification to verify *linkages* between themselves and their students during the year. Roster verification captures different teaching scenarios where multiple teachers can share instruction. Verification makes teacher analyses much more reliable and accurate.

Roster verification is completed within the EdTools platform, which is maintained by RANDA Solutions. TDOE provides EVAAS with a file that contains the approved teacher-student linkage data entered in EdTools Teacher-Student Connection accounts.

- Teacher level identification
  - Teacher Name from Tennessee Licensure Number Database (TLN DB)
  - Teacher License Number from TLN DB
- Student linking information
  - Student Last Name
  - Student First Name
  - Student Middle Initial
  - Unique Student ID (USID)
- Subjects and tests for all state TCAP Achievement, EOC, and Grade 2 Assessments
  - Semester included for EOC testing
  - Instructional Availability
  - Percent time to link
- District and school information (numbers)
- Percent of instructional responsibility (instructional time)
- Attendance flag (instructional availability)
  - F – Full
  - P – Partial
  - X – Excluded for Instructional Availability

3 Value-added analyses

As outlined in the introduction, the conceptual explanation of value-added reporting is the following:

- Growth = current achievement/current results compared to all prior achievement/prior results with achievement being measured by a quality assessment such as the Tennessee statewide assessments.

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 different entering achievement, and measurement error in the test. In Tennessee, EVAAS provides two main categories of value-added models, each comprised of district, school, and teacher reports.

- **Multivariate Response Model (MRM**) is typically used for tests given in consecutive grades, like the math and English language arts assessments in grades 3–8.
- **Univariate Response Model (URM)** is typically used when the same tests are administered to students in multiple grades, such as the EOC assessments, or when performance from previous tests is used to predict performance on another test, which might not have the same structure or subject areas, such as TCAP to ACT.

Both models offer the following advantages:

- The models include each student's testing history without imputing any test scores.
- The models can accommodate team teaching or other shared instructional practices.
- The models include multiple subjects and grades for each student to minimize the influence of measurement error.
- The models can accommodate students with different sets of testing history.
- The models can accommodate tests on different scales.

Each model is described in greater detail below.

Because the TVAAS models use multiple subjects and grades for each student, it is not necessary to make direct adjustments for students' background characteristics. In short, these adjustments are not necessary because each student serves as his or her own control. To the extent that socioeconomic/demographic influences persist over time, these influences are already represented in the student's data. As a 2004 study by The Education Trust stated, specifically with regard to the SAS EVAAS modeling:

> [I]f a student's family background, aptitude, motivation, or any other possible factor has resulted in low achievement and minimal learning growth in the past, all that is taken into account when the system calculates the teacher's contribution to student growth in the present.
>
> ***Source:*** Carey, Kevin. 2004. "The Real Value of Teachers: Using New Information about Teacher Effectiveness to Close the Achievement Gap." *Thinking K-16* 8 (1): 27.

In other words, while technically feasible, adjusting for student characteristics in sophisticated modeling approaches is not necessary from a statistical perspective, and the value-added reporting in Tennessee does not make any direct adjustments for students' socioeconomic/demographic characteristics. Through this approach, Tennessee avoids the problem of building a system that creates differential expectations for groups of students based on their backgrounds.

The value-added reporting in Tennessee is available for districts, schools, and teachers. For teachers working in multiple schools within the same district, the teacher value-added reports in the TVAAS web

application are displayed in the school for which the teacher has the largest number of full-time effective (FTE) students. FTE is explained in greater detail in Section 3.1.6.2. For teachers working in multiple districts, there is a teacher value-added report based on each individual district and displayed in that specific district's reporting in the TVAAS web application. In this instance, the teacher's evaluation composite would appear in the district for which the teacher has the largest number of FTE students.

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 *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 current 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. (See Intra-Year Approach in Section 4 on page 30.)

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

- All students with valid data are included in the analyses. Each student's testing history is included without imputing any test scores.
- By encompassing all students in the analyses, including those with missing test scores, the model provides the most realistic estimate of achievement available.
- The model minimizes the influence of measurement error inherent in academic assessments by using multiple data points of student test history.
- The model uses scores from multiple tests, including those on different scales.
- The model 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 grades and subjects simultaneously to improve precision and reliability.

Because of these advantages, the MRM is considered one of the most statistically robust and reliable approaches. The references below include studies by experts from RAND Corporation, a non-profit research organization:

- On the **choice of a complex value-added model**: McCaffrey, Daniel F., and J.R. Lockwood. 2008. "Value-Added Models: Analytic Issues." Prepared for the National Research Council and the National Academy of Education, Board on Testing and Accountability Workshop on Value-Added Modeling, Nov. 13-14, 2008, Washington, DC.
- On the **advantages of the longitudinal, mixed model approach**: Lockwood, J.R. and Daniel F. McCaffrey. 2007. "Controlling for Individual Heterogeneity in Longitudinal Models, with Applications to Student Achievement." *Electronic Journal of Statistics* 1: 223-252.
- On the **insufficiency of simple value-added models**: McCaffrey, Daniel F., B. Han, and J.R. Lockwood. 2008. "From Data to Bonuses: A Case Study of the Issues Related to Awarding

Teachers Pay on the Basis of the Students' Progress." Presented at Performance Incentives: Their Growing Impact on American K-12 Education, Feb. 28-29, 2008, National Center on Performance Incentives at Vanderbilt University.

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?

Note that, during the 2015–16 school year, the State of Tennessee suspended testing in grades 3–8 for mathematics, English language arts, science, and social studies assessments, and scale scores were not available for these assessments. As a result, the historical 2016–17 TVAAS reporting in grades 5–8 for mathematics, English language arts, and science did not include 2015-16 test scores. This section explains MRM reporting available from previous years where data were not missing, and Section 3.1.7 on page 21 explains the necessary modeling adjustments to account for that year's reporting.

3.1.1 MRM at the conceptual level

An example data set with some description of possible value-added approaches might be helpful for conceptualizing how the MRM works. Assume that ten students complete a test in two different years with the results shown in Table 1. 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
Column Mean	**49.99**	**55.79**	**5.80**
Difference between Current and Previous Score Means			**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
Column Mean	**45.01**	**54.58**	**3.97**
Difference between Current and Previous Score Means			**9.57**

The MRM uses the correlation between current and previous scores in the nonmissing 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 assigning 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 in Table 2 is 5.71 when using the MRM approach to first estimate the means in each column and taking the difference. Even in a small example such as this, the estimated difference is much closer to the difference with no missing data (Table 1) than either measure obtained by the mean of the differences (3.97) or difference of the means (9.57) in Table 2. This method of estimation has been shown, on average, to outperform both 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 for Tennessee, 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 in Section 3.1.3 on page 12. The first step in the MRM is to define the scores that will be used in the model.

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 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 above 99 and below 1, NCEs can range above 100 and below 0 to preserve their equal-interval property and to avoid truncating the test scale. For example, in a typical year in Tennessee, the average maximum NCE is approximately 115, corresponding to percentile rankings above 99.0. However, for display purposes in the TVAAS web application and to avoid confusion among users with interpretation, NCEs are shown as integers from 1-99. However, truncating would create an artificial ceiling or floor, which might bias the results of the value-added measure for certain types of students forcing the gain to be close to 0 or even negative, so the actual calculations use non-truncated numbers.

The NCEs used in EVAAS analyses are based on a reference distribution of test scores in Tennessee. The *reference distribution* is the distribution of scores on a state-mandated test for all students in each year.

[1] See, for example: S. Paul Wright, "Advantages of a Multivariate Longitudinal Approach to Educational Value-Added Assessment without Imputation," Paper presented at National Evaluation Institute, 2004.

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 is set by using each individual year to create NCEs. For more on Growth Expectation, see Section 4 on page 30.

3.1.2.3 *How EVAAS uses normal curve equivalents in MRM*

There are multiple ways of creating NCEs. EVAAS uses a method that does not assume that the underlying scale is normal since experience has shown that some testing scales are not normally distributed and this 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 below show an example of a tabulated distribution of test scores from Tennessee 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 percentage ("Percent") 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, there is some percentage of students that will receive each specific score. For example, 2.2% of students received a score of 745 in Table 3. The usual convention is to consider half of that 2.2% to be "below" and half "above." Adding 1.1% (half of 2.2%) to the 39.9% who scored below the score of 745 produces the percentile rank of 41.0 in Table 3.

Table 3: Converting tabulated test scores to NCE values

Score	Frequency	Cum Freq	Percent	Cum Pct	Ptile Rank	Z	NCE
740	2,820	48,620	2.2	37.6	36.6	-0.344	42.76
742	2,942	51,562	2.3	39.9	38.8	-0.285	44.00
745	2,880	54,442	2.2	42.2	41.0	-0.226	45.23
749	2,954	57,396	2.3	44.4	43.3	-0.169	46.45
752	3,064	60,460	2.4	46.8	45.6	-0.110	47.69
755	2,982	63,442	2.3	49.1	48.0	-0.051	48.93
757	3,166	66,608	2.5	51.6	50.4	0.009	50.19

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).

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:

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

y (in the TVAAS context) is the $m \times 1$ observation vector containing test scores (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. Fixed effects are factors within the model that come from a finite population, such as all individual schools in the state of Tennessee. In the school-level model, there is a fixed effect for every school/year/subject/grade. This matrix would have a row for each of these combinations.

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. In contrast to fixed effects, random effects do not come from a fixed population but rather can be thought of as a random sample coming from a large population where not all individuals in that population are known. This is more appropriate for the teacher model for many reasons: not all teachers are included (e.g., small class sizes), new teachers start each year while others leave each year, etc. As such, teachers are treated as random factors in this model.

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:[2]

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

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

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

[2] Sanders, William L., Arnold M. Saxton, and Sandra P. Horn. 1997. "The Tennessee Value-Added Assessment System: A Quantitative, Outcomes-Based Approach to Educational Assessment." In Grading Teachers, Grading Schools, ed. Jason Millman, 137-162. Thousand Oaks, CA: Sage Publications.

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.

$$E(K^T\beta) = K^T b \tag{5}$$

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

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

$$E(v|u) = u \tag{7}$$

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

   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).

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

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

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.

This Section describes the technical details specifically around the MRM approach. However, many more details describing the linear mixed model can be found in various statistical texts. [3]

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. Since MRMs are generally run with each school uniquely defined across districts, there is no need to include district in the 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"

[3] See, for example, Charles E. McCulloch, Shayle R. Searle, and John M. Neuhaus, *Generalized, Linear, and Mixed Models* (Hoboken, NJ: Wiley, 2008).

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 uses 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 multiple years of data 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. Each level of analysis will utilize the values that are found within that 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 (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 fed students into the school, grade, subject, and year in question. Prior year schools are not utilized if they are feeding students in very small amounts (less than 5) 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 tested 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 who 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 includes (1) cumulative gains across grades (for each subject and year), and (2) up to 3-year average gains (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*

As a protection to teachers, the teacher estimates use a more conservative statistical process to lessen the likelihood of misclassification. Each teacher effect is assumed to be the state average in a specific year, subject, and grade until the weight of evidence pulls the teacher effect either above or below that state average. Furthermore, the teacher model is a "layered" model, which means that:

- The current and previous teacher effects are incorporated.
- Each teacher estimate takes into account all the students' testing data over the years.
- The percentage of instructional responsibility (instructional time) the teacher has for each student is used.

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

For reasons described when introducing random effects, the MRM treats 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 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 t 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:

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

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 those teachers. $w_{ijk^*l^*t}$ 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 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 mean gain. That is, the interpretation of a teacher effect in this teacher model is expressed as a deviation from the average gain for the state in a given year, subject, and grade.

Table 4 below illustrates how the Z matrix is encoded for three students who have three different scenarios of teachers during grades 3, 4, and 5 in two subjects, math (M) and English language arts (R). In Tennessee, this matrix would include science, but this illustrates how it is encoded.

Tommy's teachers represent the conventional scenario: Tommy is taught by a single teacher in both subjects each year (teachers Abbot, Card, and East in grades 3, 4, and 5, respectively). Notice that in Tommy's Z matrix rows for grade 4, there are ones (representing the presence of a teacher effect) not only for fourth grade teacher Card but also for third grade teacher Abbot. This is how the "layering" is encoded. Similarly, in the grade 5 rows, there are ones for grade 5 teacher East, grade 4 teacher Card, and grade 3 teacher Abbot.

Susan is taught by two different teachers in grade 3, teacher Abbot for math and, teacher Banks for English language arts. In grade 4, Susan had teacher Card for English language arts. For some reason, in grade 4 no teacher claimed Susan for math even though Susan had a grade 4 math test score. This score can still be included in the analysis by entering zeros into the Susan's Z matrix rows for grade 4 math. In grade 5, on the other hand, Susan had no test score in English language arts. 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 5 English language arts, there can be no corresponding row in Z.

Eric's scenario illustrates team teaching. In grade 3 English language arts, Eric received an equal amount of instruction from both teachers Abbot and Banks. The entries in the Z matrix indicate each teacher's contribution, 0.5 for each teacher. In grade 5 math, however, while Eric was taught by both teachers East and Farr, they did not make an equal contribution. Teacher East claimed 80% responsibility and teacher Farr claimed 20%. If a student is claimed at more than 100%, then the model will adjust the percentage of instructional responsibility of each teacher proportional to the amount claimed such that the overall percentage is 100%. The model does not make adjustments to students who are claimed at less than 100%. In other words, if teacher Abbot claimed Eric at 100% for math and teacher Card claimed Eric at 50% for math, then teacher Abbot's instructional responsibility for Eric would be weighted at 100/150 and teacher Card's would be weighted at 50/150. If a student is claimed at less 100%, then the percentages submitted are used in this model.

Teacher effect estimates are obtained by shrinkage estimation, technically known as best linear unbiased prediction or as empirical Bayesian estimation. This is a characteristic of random effects from a mixed model and means that *a priori* a teacher is considered "average" (with a teacher effect of zero) until there is sufficient student data to indicate otherwise. Zero represents the statewide average teacher effect in this case. 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).

The teacher model provides estimated mean gains for each subject and grade. These quantities can be described by linear combinations of the fixed and random effects and are found using the equations mentioned above. In each year's analysis, multiple years of teacher value-added measures are

calculated within the models; however, only the teacher gains from the current year are used since the re-estimated prior year measures are no longer being used.

Table 4: Encoding the Z matrix in a typical testing scenario

			Third Grade				Fourth Grade				Fifth Grade			
			Abbot		Banks		Card		Dupont		East		Farr	
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
		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

3.1.4 Where the MRM is used in Tennessee

Typically, the MRM is used in math, English language arts, and science to provide value-added measures at the district, school, and teacher level. Note that there is no historical reporting based on the 2015-16 school year for these assessments since scale scores were not available. Section 3.1.7 on page 21 provides more details about how the 2016-17 reporting was modified to accommodate these missing scores.

The MRM methodology provides estimated measures of progress for up to three years in each subject/grade/year for district, school and teacher analyses provided that the minimum student requirements are met (details in Section 3.1.6 on page 20). For each subject, growth measures might be available across grades, years, and combined years and grades.

At the teacher level, value-added measures for each subject/grade/year are computed (and displayed on the TVAAS web application available at https://tvaas.sas.com/).

More information about teacher composite measures that use teacher data from up to three years can be found in Section 6 on page 36.

3.1.5 Students included in the analysis

All students' scores are included in these analyses if the scores can be used and do not meet any criteria for exclusion outlined below or in Section 8 on page 48. In other words, every student's math, English language arts, and science results for the student's cohort are incorporated into the models.

Business rules for excluding scores are as follows. First-time EL test takers who have no prior testing history will not be included in the analysis the first time that they test. These students will be included in future years if they have prior scores that can be used in the analysis.

The analysis also excludes all scores that do not have an "Overall RI Status" of zero.

A student score could be excluded if it is considered an "outlier" in context with all the other scores in a reference group of scores from an individual student. In other words, is the score "significantly different" from the other scores as indicated by a statistical analysis that compares each score to the other scores? There are different business rules for the low outlier scores and the high outlier scores, and this approach is more conservative when removing a very high achieving score. In other words, a lower score would be considered an outlier before a higher score would be considered an outlier. More details are provided in Section 8 on page 48.

3.1.5.1 *District and school measures*

3.1.5.1.1 Overall measures of student growth for districts and schools

The analyses for schools and districts include all applicable student scores from math, English language arts, and science tests from the cohort of students testing in the most recent three years.

3.1.5.1.2 Subgroup measures of student growth for districts and schools

Tennessee uses subgroup-level value-added measures in their federal accountability system. This section describes what students are included in each analysis. In each subgroup value-added computation, the expectation of growth is defined the same as in the overall students' analysis. In other words, the expectation of growth is based on all students. Furthermore, the estimated covariance parameters are used from the overall students' analysis when calculating the value-added measures. These measures are provided using the TCAP subjects with a composite across math in grades 4–8 and English language arts in grades 4–8.

3.1.5.1.2.1 Subgroup: Economically disadvantaged district- and school-level analysis

The economically disadvantaged student analysis pertains only to those students with a code "A" flag for economically disadvantaged. Value-added measures are calculated for this subset of students for each district and school that meet the minimum requirements of student data.

3.1.5.1.2.2 Subgroup: Students with disabilities district- and school-level analysis

The students with disabilities analysis pertains only to those students who are denoted as students with disabilities as recorded by the special ED flag as "Yes." Value-added measures are calculated for this subset of students for each district and school that meet the minimum requirements of student data.

3.1.5.1.2.3 Subgroup: EL students district- and school-level analysis

The EL students' analysis pertains to those students who are denoted as English Learner students or who are classified as EL or T1 – T4. Value-added measures are calculated for this subset of students for each district and school that meet the minimum requirements of student data.

3.1.5.1.2.4 Subgroup: Black/Hispanic/Native American students district- and school-level analysis

The students identified as Black/Hispanic/Native American analysis pertains only to those students who are denoted with a race category of Black or African American, Hispanic/Latino, or Native American or Other Pacific Islander. Value-added measures are calculated for this subset of students for each district and school that meet the minimum requirements of student data.

3.1.5.1.2.5 Super Subgroup: Economically disadvantaged, students with disabilities, EL students, or Black/Hispanic/Native American students district- and school-level analysis

One additional subgroup value-added measure is created by combining the four subgroups together that are described above into a "super subgroup." Value-added measures are calculated for this subset of students for each district and school that meet the minimum requirements of student data.

3.1.5.2 *Teacher measures*

The teacher value-added reports use all available test scores for each individual student linked to a teacher through roster verification, unless a student or a student test score meet certain criteria for exclusion.

Students are excluded from the teacher analysis if the students have an attendance flag in the student-teacher linkages of P or X, meaning they were partially claimed or excluded for instructional availability. A student is excluded from the teacher analysis if he or she does not have any prior test scores in the same subject in any prior year. In both cases, the students' scores are still included in the model, but they are not connected to any individual teachers.

3.1.6 Minimum number of students for reporting

3.1.6.1 *District and school level*

To ensure that estimates are reliable, the minimum number of students required to report an estimated mean NCE score for a school or district in a specific subject/grade/year is six.

To report an estimated NCE gain for a school or district in a specific subject/grade/year, there are additional requirements:

- There must be at least six students who are associated with the school or district in that subject/grade/year.
- There is at least one student at the school or district who has a "simple gain," which is based on a valid test score in the current year/grade as well as the prior year/grade in the same subject.
- Of those students who are associated with the school or district in the current year/grade, there must be at least five students that have come from any single school for that prior school to be used in the gain calculation.

3.1.6.2 *Teacher*

The teacher value-added model includes teachers who are linked to at least six students with a valid test score in the same subject and grade. To clarify, this means that the teachers are included in the analysis, even if they do not receive a report due to the other requirements. In other words, this requirement

does not consider the percentage of instructional time that the teacher spends with each student in a specific subject/grade.

However, to receive a teacher value-added report for a particular year, subject, and grade, there are two additional requirements. First, a teacher must have at least six full-time equivalent (FTE) students in a specific subject/grade/year. The teacher's number of FTE students is based on the number of students linked to that teacher and the percentage of instructional time the teacher has for each student. For example, if a teacher taught 10 students for 50% of their instructional time, then the teacher's FTE number of students would be five, and the teacher would not receive a teacher value-added report. If another teacher taught 12 students for 50% of their instructional time, that teacher would have six FTE students, and that teacher would receive a teacher value-added report. The instructional time attribution is obtained from the student-teacher linkage data. This information is in the files sent to EVAAS described in Section 2 on page 3. As the second requirement, the teacher must be linked to at least five students with prior test score data in the same subject, and the test data might come from any prior grade so long as they are part of the student's regular cohort (meaning, if a student repeats a grade, then the prior test data would not apply as the student has started a new cohort). One of these five students must have a "simple gain," meaning the same subject prior test score must come from the immediate prior year and prior grade. Students are linked to a teacher based on the subject area taught and the assessment taken.

3.1.7 Modeling adjustments to 2016-17 growth measures in grades 5–8 to accommodate missing 2015-16 data for grades 3–8

3.1.7.1 *Overview*

During the 2015–16 school year, the State of Tennessee suspended testing in grades 3–8 for mathematics, English language arts, science, and social studies assessments, and scale scores were not available for these assessments. As a result, the historical 2016-17 TVAAS reporting in grades 5–8 for mathematics, English language arts, and science does not include 2015-16 test scores.

Because statewide testing begins in grade 3, it is not possible to provide a growth measure for grade 4 without the 2015-16 test scores in grade 3. Although a small subset of districts has grade 2 data from 2014-15, TDOE will have a consistent approach across the state as grade 4 TVAAS is typically available to all districts in the state through the gain-based model for measuring growth. The exclusion of grade 4 growth measures applies to district, school and teacher reporting for 2016-17.

To conceptualize what the 2016-17 growth measures mean for districts and schools in grades 5–8, Table 5 below provides the average achievement level for the students testing at a sample school. As a cohort of students moves from one grade to the next, their achievement level can be tracked along a diagonal line. For example, Table 5 shows that the achievement level of grade 5 students in year 2 is 25 NCEs and then changes to 36 NCEs when this cohort of students is in grade 6 in year 3.

Table 5: Average Achievement in NCEs by Grade and Year for Sample School

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Year 1	13	14	15	16	17	18
Year 2	23	24	25	26	27	28
Year 3	33	34	35	36	37	38

In the computationally ideal situation where all students are present in all three years and students never change schools, the calculation of gains is straightforward. To calculate the gain for grade 6 in year 3, it would be the achievement level for grade 6 in year 3 minus the achievement level for grade 5 in year 2. That would be 36 NCEs minus 25 NCEs, or 11 NCEs.

In reality (not the computationally ideal situation described above), the MRM calculates means by taking into account missing student scores and allowing for students who move between schools. The achievement level reported for grade 6 in year 3 is often a weighted average based on the number of students coming from the sample school's feeder schools. This is relevant for the lowest grade in a school, often grade 6, since there is no mean *at that school* for the previous grade and year.

In either instance (the computationally ideal situation or the weighted average based on feeder schools), there is data available to calculate single-year gains.

If there is no year 2 data, it is not possible to calculate *single-year gain* for grade 6 in year 3. It is possible, however, to calculate a *cumulative gain* based on the change in achievement from grade 4 in year 1 to grade 6 in year 3. This would be 36 NCEs minus 14 NCEs, or 22 NCEs.

To determine the feasibility of this approach, the cumulative gain could be compared to the sum of the single-year gains based on a model with year 2 data. This would be (36 NCEs – 25 NCEs) + (25 NCEs – 14 NCEs), which would be 11 NCEs + 11 NCEs, or 22 NCEs. The ideal case is that the cumulative gain and the sum of the single-year gains are the same. In practice, however, they might differ due to lack of information about weighting feeder schools and missing student data. This simulation research described below provides insight as to how this might differ with actual Tennessee assessment data.

3.1.7.2 *Research on Missing Year Data*

To confirm that the cumulative gain is an appropriate measure to provide to districts and schools in 2016-17, TDOE asked EVAAS to conduct simulation research using prior years' data. To approximate the 2016-17 reporting, which is missing the immediate prior year of data, the simulation research compared a sum of single-year 2013-14 and 2014-15 MRM growth measures (which did not have a year of data missing) to an MRM growth measure spanning 2012-13 to 2014-15 (which excluded the immediate prior year of data, the 2013-14 test scores). This scenario was evaluated for district, school, and teacher models to determine the correlation between 2014-15 growth measures as calculated and 2014-15 growth measures excluding data from the prior year. The results of the simulation research comparisons are summarized in Table 6 below.

The correlation reports the strength of the relationship between variables with +1 indicating a perfect positive relationship (positive meaning, when one variable changes, the other variable changes in a similar way) and -1 indicating a perfect negative relationship (meaning, when one variable changes, the other variable changes in an opposite way). While a precise definition varies, a typical interpretation of the correlation is that a weak relationship is between 0.10 and 0.30, a moderate relationship is between 0.30 and 0.50, and a strong relationship is above 0.50.[4]

The district and school models show that the results for growth measures in 2014-15 with and without the prior year data are very similar, with a correlation above 0.99 in both the district and school results. Another way to assess the practical implications of the relationship between the two models is to note how many growth indices stayed or changed their level categorization between the two models. Of the 1,656 growth indices in the district comparison, 1,550 (93.6%) stayed the same level, 41 (2.5%) moved up one level, 61 (3.7%) moved down one level, 2 (0.1%) moved up two or more levels, and 2 (0.1%)

[4] Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

moved down two or more levels. Of the 7,637 growth indices in the school comparison, 6,964 (91.2%) stayed the same level, 285 (3.7%) moved up one level, 373 (4.9%) moved down one level, 7 (0.1%) moved up two or more levels, and 8 (0.1%) moved down two or more levels. In each comparison, a fairly even percentage growth indices moved up or down a level (or up or down two or more levels).

The teacher analyses provide a strong correlation in growth measures between the two models, comparing 2014-15 growth measures with and without the prior year data available. The correlation between the models is 0.80. Again, another way to assess the practical implications of the relationship between the two models is to note how many growth indices stayed or changed their level categorization between the two models. Of the 16,055 growth indices in the teacher comparison, 9,153 (57.0%) stayed the same level, 2,421 (15.1%) moved up one level, 2,380 (14.8%) moved down one level, 1,204 (7.5%) moved up two or more levels, and 897 (5.6%) moved down two or more levels.

Table 6: Comparing Cumulative Gain MRM With and Without Missing Year of Data for District, School, and Teacher Growth Indices by Subject/Grade: Change in Level Categorization

Value-added model	Correlation (r)	Level stayed the same (%)	Moved up 1 or more levels (%)	Moved down 1 or more levels (%)
District	.99	93.6	2.6	3.8
School	.99	91.2	3.8	5.0
Teacher	.80	57.0	22.6	20.4

As another point for interpretation, it is worth reiterating that, because the Missing Year MRM provides growth measures spanning two years of schooling, the growth measure for grades where students transition from one school to another will then include growth from the feeder school(s) as well as the receiver school. For example, in these models, a middle school with grades 6–8 could receive a growth measure for sixth grade based on the students' growth in sixth grade as well as their growth from the feeder elementary school(s) in fifth grade. In other words, it is not possible to completely parse out the individual contribution of the middle school in sixth grade apart from those from the elementary school(s) in fifth grade because of the missing year of test scores.

For the district growth measures and for the non-transition grades, the cumulative growth measure would not have the same limitation. The district growth measures are still representative of growth within the specific district, and the non-transition grades for the school are still representative of growth within the specific school. Thus, we still find a strong correlation between the growth measures with and without prior year data despite this limitation of data from the transition year to a new school.

3.2 Univariate Response Model (URM)

Tests that are not necessarily administered to students in consecutive years, like the EOC tests, require a different modeling approach from the MRM, and this modeling approach is called the univariate response model (URM). This model is also used when previous test performance is used to predict another test performance, such as the TCAP Social Studies in grades 5–8 or ACT. 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, such as the case with TCAP grade 3 reporting using the Grade 2 Assessments as historical predictors.

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

- The model does not require students to have all predictors or the same set of predictors if a student has at least three prior test scores in any subject/grade.
- The model 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.
- The model uses scores from multiple tests, including those on different scales.
- The model accommodates teaching scenarios where more than one teacher has responsibility for a student's learning in a specific subject/grade/year.

In Tennessee, URM value-added reporting is available for Grade 2 Assessments (for teacher reports from previous years), grade 3 assessments, and all EOC assessments at the district, school, and teacher levels.

The availability of grade 2 and 3 measures depends on past and current participation in the optional early grades assessments. For this year's reporting, grade 3 data will be available in districts that administered the optional Grade 2 Assessment in 2017-18. As assessments prior to second grade are no longer administered, second grade TVAAS data is not available following the 2017-18 school year. Grade 3 data will continue to be available for districts that continue to administer the optional second grade assessment.

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 Biology in a given year. Those students are connected to their prior testing history (across grades, subjects, and years), and the relationship between the observed Biology 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 example, it might be that prior science tests will have a greater relationship with Biology than prior English language arts 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. 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 approach is described briefly below with more details following.

- 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.
- 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).

Algebraically, the model can be represented as follows for the i^{th} student when there is no team teaching.

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

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). Similar to what was explained in the MRM section, if a student is claimed at more than 100%, then the model will adjust the percentage of instructional responsibility of each teacher proportional to the amount claimed such that the overall percentage is 100%. The model does not make adjustments to students who are claimed at less than 100%.

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 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:

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

Two difficulties must be addressed 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:

$$C = \begin{bmatrix} C_{yy} & C_{yx} \\ C_{xy} & C_{xx} \end{bmatrix} \quad (16)$$

Note that C in equation (16) is not the same as C in equation (4). This matrix is estimated using an Expectation Maximization (EM) algorithm for estimating covariance matrices in the presence of missing data such as the one provided in the SAS/STAT® MI Procedure, 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:

$$\hat{\beta} = C_{xx}^{-1} c_{xy} \quad (17)$$

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 the parameters are defined such 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 means. The teacher 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 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.

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

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) 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 English language arts, 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:

$$y_i = \gamma_0 + \gamma_1 \hat{y}_i + \alpha_j + \epsilon_i \quad (19)$$

In the URM model, the effects ($\boldsymbol{\alpha_j}$) are considered 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.

3.2.3 Where the URM is used in Tennessee

In Tennessee, URM value-added reporting is available for grade 3 assessments, all EOC assessments for districts, schools, and teachers, and ACT for districts and schools.

The URM methodology provides estimated measures of progress for up to three years in each subject/grade/year for district, school, and teacher analyses provided that the minimum student requirements are met (details in Section 3.2.5 below). For each subject, growth measures might be available across grades, years, and combined years and grades.

3.2.4 Students included in the analysis

3.2.4.1 *Overall measures of student growth for districts, schools and teachers*

All students' scores are included in these analyses if the scores can be used and do not meet any criteria for exclusion outlined below or in Section 8 on page 48.

Business rules for excluding scores are as follows. First-time EL test takers who have no prior testing history will not be included in the analysis the first time that they test. These students will be included in future years if they have prior scores that can be used in the analysis.

The analysis also excludes all scores that do not have an "Overall RI Status" of zero, which indicates that no reports of irregularity were submitted for issues such as test misadministration.

A student score could be excluded if it is considered an "outlier" in context with all the other scores in a reference group of scores from an individual student. In other words, is the score "significantly different" from the other scores, as indicated by a statistical analysis that compares each score to the other scores? There are different business rules for the low outlier slier scores, and this approach is more conservative when removing a very high achieving score. In other words, a lower score would be considered an outlier before a higher score would be considered an outlier. More details are provided in Section 8.

For the teacher analysis, students are excluded if they have a P or an X value entered for instructional availability in the student-teacher linkages data.

Furthermore, for a student's score to be used in the district- or school-level analysis for a particular subject/grade/year, the student must have at least three valid predictor scores that can be used in the analysis, all of which cannot be deemed outliers. These scores can be from any year, subject, and grade that are used in the analysis. It will include subjects other than the subject being predicted. The required three predictor scores are needed to sufficiently dampen the error of measurement in the tests to provide a reliable measure. If a student does not meet the three-score minimum, then that student is excluded from the analyses. Not all students have to have the same three prior test scores; they only have to have some subset of three that were used in the analysis.

3.2.4.2 *Subgroup measures of student growth for districts and schools*

Tennessee uses subgroup-level value-added measures in their federal accountability system. This section describes which students are included in each analysis. In each subgroup value-added computation, the expectation of growth is defined the same as in the overall students' analysis. Therefore, the expectation of growth is based on all students. Furthermore, the estimated covariance parameters are used from the overall students' analysis when calculating the value-added measures. These measures are provided using the EOC subjects with a composite across Algebra I, II, Integrated Math I, II, III, and Geometry as well as a composite across English I, II, and III. More details about how these subgroup measures of growth are combined with those from MRM reporting are available in Section 6.2.8.

3.2.4.2.1 Subgroup: Economically disadvantaged district and school level analysis

The economically disadvantaged student analysis pertains only to those students with a code "A" flag for economically disadvantaged. Value-added measures are calculated for this subset of students for each district and school that meet the minimum requirements of student data.

3.2.4.2.2 Subgroup: Students with disabilities district and school level analysis

The students with disabilities analysis pertains only to those students who are denoted as students with disabilities as recorded by the special ED flag as "Yes." Value-added measures are calculated for this subset of students for each district and school that meet the minimum requirements of student data.

3.2.4.2.3 Subgroup: EL students district and school level analysis

The EL students' analysis pertains to those students who are denoted as English Learner students or who are classified as either EL or T1 – T4. Value-added measures are calculated for this subset of students for each district and school that meet the minimum requirements of student data.

3.2.4.2.4 Subgroup: Black/Hispanic/Native American students district- and school-level analysis

The students identified as Black/Hispanic/Native American analysis pertains only to those students who are denoted with a race category of Black or African American, Hispanic/Latino, or Native American or Other Pacific Islander. Value-added measures are calculated for this subset of students for each district and school that meet the minimum requirements of student data.

3.2.4.2.5 Super Subgroup: Economically disadvantaged, students with disabilities, EL students, or Black/Hispanic/Native American Students district- and school-level analysis

One additional subgroup level value-added measure is created by combining the four subgroups together that are described above into a "super subgroup." Value-added measures are calculated for this subset of students for each district and school that meet the minimum requirements of student data.

3.2.5 Minimum number of students for reporting

To receive a report, a district or school must have at least 10 students in that year, subject, and grade that have the required three prior test scores needed to obtain a predicted score in that year, subject, and grade and have met all other requirements to be included.

For teacher reporting, there must be 10 students meeting criteria for inclusion in that year, subject, and grade that have the required three prior test scores needed to obtain a predicted score in that year, subject, and grade. Again, to receive a teacher value-added report for a particular year, subject, and grade, a teacher must have at least six Full Time Equivalent (FTE) students in a specific subject/grade/year as described in Section 3.1.6.2.

3.2.6 Use of ACT data in the analysis

The TVAAS reporting for ACT uses students' end-of-grade predictors through grade 8. This allows TVAAS reporting for ACT to assess students' growth over the course of their high school career until they take the ACT.

For the purposes of TVAAS reporting, the business rules for ACT are slightly different from those for EOC. Because students might take the ACT several times throughout their high school career, there might be several possible scores to use in TVAAS reporting. In contrast, students typically take the EOC assessment at the same point in the course (unless they do not pass and need to re-take the test). For a more equitable comparison of students' schooling experiences across a similar point in time, TVAAS district and school growth measures for ACT use the "junior day" data, meaning the test score obtained during students' junior year of high school.

It is also important to note that multiple high school courses can prepare students for each ACT subject area. For that reason, district- and school-level TVAAS reporting for ACT are available but teacher-level is not.

4 Growth expectation

The simple definition of growth was described in the introduction as follows:

- Growth = current achievement/current results compared to all prior achievement/prior results with achievement being measured by a quality assessment such as the Tennessee statewide tests

Typically, the "expected" growth is 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 precise definition of "expected growth" varies by model, and this section provides more details.

As a reminder, during the 2015-16 school year, not all districts administered Part II of mathematics, English language arts, science, and social studies assessments in grades 3–8. As a result, scale scores will not be available for these assessments, and TVAAS data will not be provided in these grades and subjects. These grades and subjects historically received reporting based on the intra-year approach described below, so the MRM is still included in the description as a reference for reporting from prior years. For the 2016-17 reporting, the concept is similar with the exception that it is based on a period from 2014-15 to 2016-17.

4.1 Intra-year approach

4.1.1 Description

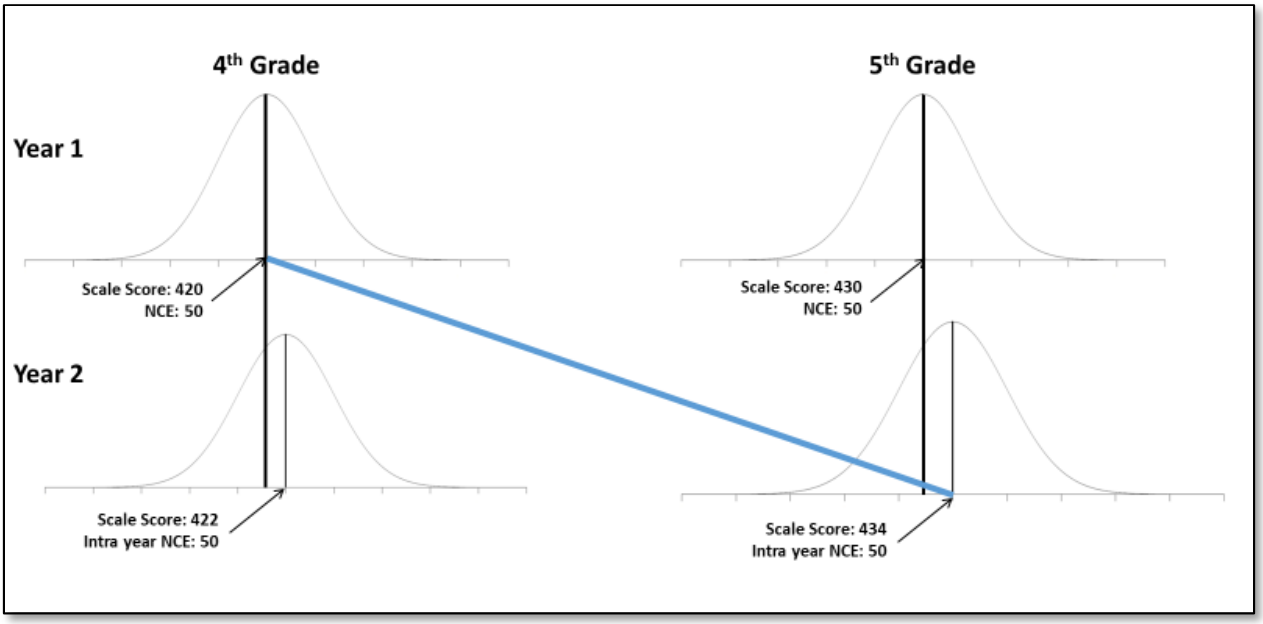
- This approach has always been used in Tennessee with the URM reporting and was used for the first time for the 2014-15 testing with the MRM reporting.
- The actual definitions in each model are slightly different, but the concept can be considered as the average amount of progress seen across the state 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 might 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. For example, if students' achievement was at the 50th NCE in 2017 grade 4 math, based on the 2017 grade 4 math statewide distribution of student achievement, and their achievement is at the 50th NCE in 2018 grade 5 math, based on the 2018 grade 5 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. However, it should be noted that there is not a set distribution of the value-added measures and being centered on the growth expectation does not mean half of the measures would be in the positive levels and half would be in the negative levels since many value-added measures are indistinguishable from the expectation when considering the statistical certainly around that measure. More is explained about this in Section 5.

4.1.2 Illustrated example

Figure 1 below provides a simplified example of how growth is calculated with an intra-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 figure shows how the gain is calculated for a group of fourth grade students in year 1 as they become fifth grade students in year 2. In year 1, our fourth-grade students score, on average, 420 scale score points on the test, which corresponds to the 50th NCE (similar to the 50th percentile). In year 2, the students score, on average, 434 scale score points on the test, which corresponds to a 50th NCE *based on the grade five distribution of scores in year 2*. The fifth-grade distribution of scale scores in year 2 was higher than the fifth-grade distribution of scale scores in year 1, 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 be to maintain their position at the 50th NCE for fourth grade in year 1 as they become fifth grade students in year 2. The growth measure for these students is year 2 NCE – year 1 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.

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.

Figure 1: Intra-Year Approach Example



4.2 Base-year approach (used in prior years' value-added measures)

4.2.1 Description

In years prior to the 2014-2015 school year, the MRM value-added models used a "base-year approach." This means that the 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 result, prior years' value-added measures, which are incorporated in multi-year trends on the value-added reports, use the base-year approach, and this section provides an overview of that how the growth expectation is derived for those measures.

As a simplified example with 2013 as the base year, if students' achievement was at the 50th NCE in 2013 grade 4 math based on the 2013 grade 4 math scale score distribution and at the 52nd NCE in 2014 grade 5 based on the 2013 grade 5 math scale score distribution, then their estimated mean gain is 2 NCEs.

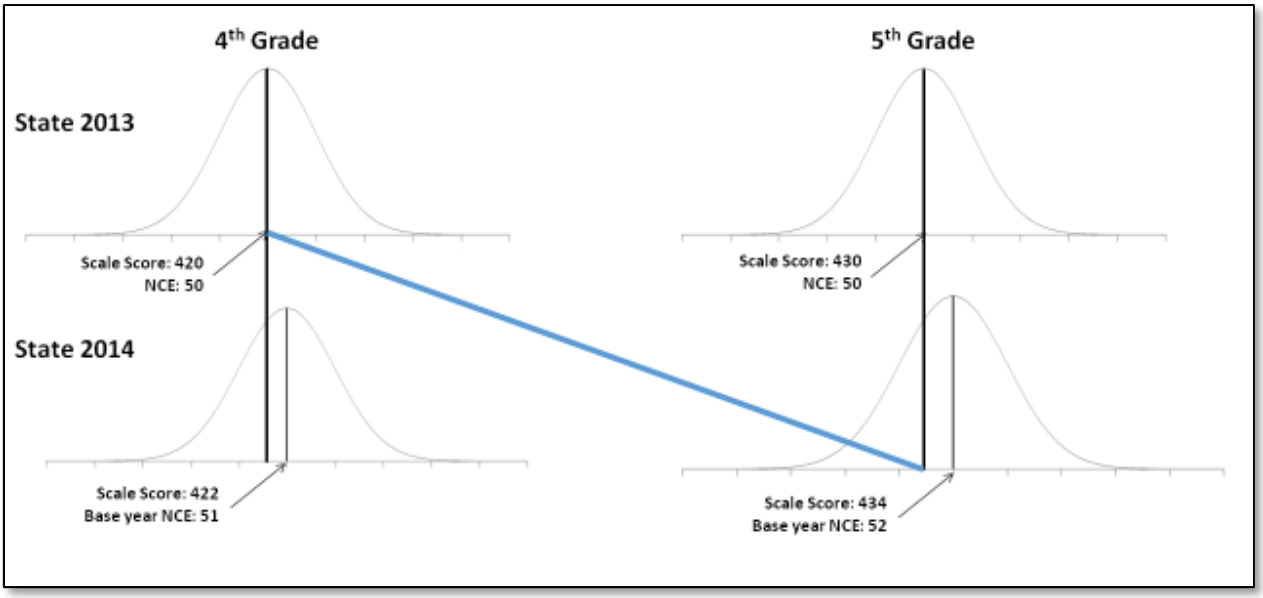
The key feature is that, in theory, all educational entities could exceed or fall short of the growth expectation (or standard) in a particular subject/grade/year, and the distribution of entities that are considered above or below could change over time.

4.2.2 Illustrated example

Figure 2 below 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 2013, and the graphic shows how the gain is calculated for a group of 2013 fourth grade students as they become 2014 fifth-grade students. In 2013, our fourth-grade students score, on average, 420 scale score points on the test, which corresponds to the 50th NCE (similar to the 50th percentile). In 2014, the students score, on average, 434 scale score points on the test, which corresponds to a 52nd NCE based on the 2013 fifth grade distribution of scores. The 2014 fifth grade distribution of scale scores was higher than the 2013 fifth grade 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 be to maintain their position at the 50th NCE in 2013 fourth grade as they become 2014 fifth-grade students. The growth measure for these students is 2014 NCE – 2013 NCE, which would be 52 – 50 = 2. Similarly, if a group of students started out at the 35th NCE in 2013 fourth grade and then moved their position to the 37th NCE in 2014 fifth grade, they would have a gain of two NCEs as well.

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 simple illustration provides the basic concept.

Figure 2: Base-Year Approach Example



4.3 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 changes 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 intra-year approach in the MRM, the scales from one year to the next can be completely different from one another. This method converts any scale to a relative position and can be used through an assessment change.

Over the past 20 years, TVAAS reporting has accommodated different changes in testing regimes and used several tests for the MRM without a break in reporting, such as the Comprehensive Test of Basic Skills/4 (CTBS/4), TerraNova, Tennessee Comprehensive Assessment Program Criterion Referenced Test (TCAP-CRT), and Tennessee Comprehensive Assessment Program Achievement (TCAP).

5 Using standard errors to create levels of certainty and define effectiveness

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

5.1 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 about 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 use 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. As described in Section 3 on page 7, there are minimum requirements of students per tested subject/grade/year depending on the model, which are relatively small.

The standard error also takes into account that, even among teachers with the same number of students, teachers might 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 into value-added reporting.

5.2 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. In the reporting, there is a need to display the values used to determine these categories. This value is typically referred to as the growth index and is simply the value-added measure divided by its standard error. **Since the expectation of growth is zero, this measures the certainty about the difference of a growth measure to zero.**

The 2017 Value Added reports for districts, schools, and teachers are color-coded as follows.

Value Added Color	District and School Growth Measure Compared to the Growth Standard	Index*	Interpretation
Level 5 Most Effective	At least 2 standard errors above	2.00 or greater	Significant evidence that students exceeded the Growth Standard.
Level 4 Above Average Effectiveness	Between 1 and 2 standard errors above	Between 1.00 and 2.00	Moderate evidence that students exceeded the Growth Standard.
Level 3 – Average Effectiveness	Between 1 standard error above and 1 standard error below	Between -1.00 and 1.00	Evidence that students met the Growth Standard.
Level 2 – Approaching Average Effectiveness	Between 1 and 2 standard errors below	Between -2.00 and -1.00	Moderate evidence that students did not meet the Growth Standard.
Level 1 Least Effective	More than 2 standard errors below	Less than -2.00	Significant evidence that students did not meet the Growth Standard.

NOTE: When an index falls exactly on the boundary between two colors, the higher growth color is assigned.

***These rules for effectiveness levels and growth colors apply to all index values in the district, school, and teacher reports.**

The distribution of these categories can vary by year/subject/grade. There are many reasons this is possible, but overall, these categories are based on the amount of evidence that shows whether students make more or less than the expected progress.

5.3 Rounding and truncating rules

As described in the previous section, the effectiveness categories are based on the value of the growth index. In determining the growth index, rounding and truncating rules are applied only in the final step of the calculation. Thus, the calculation of the growth index uses unrounded values for the value-added measures and standard errors. After the growth index has been created but before the categories are determined, the index values are rounded or truncated by taking the maximum value of the rounded or truncated index value out to two decimal places. This business rule yields the highest category of effectiveness given any type of rounding or truncating situation. For example, if the index score was a 1.995, then rounding would provide a higher category. If the score was a -2.005, then truncating would provide a higher category. In practical terms, this only impacts a very small number of measures.

Also, when value-added measures are combined to form composites, as described in the next section, the rounding or truncating occurs after the final index is calculated for that combined measure.

6 TVAAS composite calculations

6.1 Teacher evaluation composites

Teachers might receive evaluation composites based on their individual TVAAS value-added reporting, and teachers with a 2017-18 TVAAS teacher value-added measure are eligible to receive one of these composites. The composite that yields the highest "overall level of effectiveness" when combined with other components of the evaluation will be used.

For the 2017-18 reporting year, there are up to three evaluation composites available for each teacher. The first is a **single-year evaluation composite** comprised solely of value-added measures from the current year reporting (i.e., 2017-18). The second is a **multi-year evaluation composite** that includes up to three years' reporting (i.e., 2015-16, 2016-17, and 2017-18) together at 35%, where 2015-16 and 2016-17 is weighted at 25% and 2017-18 at 10%. This composite is available for any teachers with 2017-18 TVAAS data and 2015-16 and/or 2016-17 TVAAS data. The third option is a **multi-year evaluation composite without 2015-16 data**, and this composite excludes any 2015-16 data and weighs 2016-17 data and 2017-18 data at 10% each. This adjusted multi-year option is available only for teachers who received 2015-16 teacher TVAAS reports. This section provides examples of how composite type is calculated.

Note that the value-added measures within the composite for a given year will be weighted according to the number of Full-Time Equivalent (FTE) students associated with each value-added measure.

Also, the evaluation composite will include all available value-added measures within the years defined above for a teacher. Through 2014-15, only value-added measures from subjects taught in the current year (and from the same model) were included in the evaluation composite.

Details for how these composites are incorporated in each available weighting option for educator evaluation are available in TDOE's Detailed teacher evaluation guidance for 2017-18 document.

6.1.1 Sample calculation of teacher evaluation composite

The table below provides sample value-added measures for a teacher to illustrate how the evaluation composite would be calculated.

Table 7: Sample value-added measures for a teacher

Year	Subject	Number of FTE Students	Value-Added Measure	Standard Error	Index
2016	Biology I	25	3.47	1.60	2.17
2016	Algebra II	100	3.50	1.50	2.33
2017	Algebra I	50	0.50	1.40	0.36
2017	Algebra II	50	4.50	1.60	2.81
2018	Geometry	50	-0.30	1.20	-0.25
2018	Algebra II	50	3.80	1.50	2.53
2018	Algebra I	25	15.50	5.50	2.82

Note that teacher evaluation composites could contain more than one scale since the various EOC assessments are in different scales. Therefore, the value-added measures cannot simply be averaged

across the seven different subject/grade/years for this sample teacher's evaluation composite. An index value can be used to combine them.

The index is standardized (unit-less) or in terms of the standard errors away from zero. This makes it possible to combine across subjects and grades. By definition, according to standard statistical theory, this standardized statistic has a standard error of 1.[5] The index is calculated for each teacher's value-added measure by dividing the value-added measure by its standard error. The index is reported in the final column of Table 7. As a reminder from earlier sections, the model produces a value-added measure and standard error for each year/subject/grade possible for a teacher. These two values are used to see whether there is statistical evidence that the value-added measure is different from the expectation of growth, which is zero.

6.1.2 Calculation of the single-year evaluation composite

To calculate the single-year evaluation composite, the first step is to average the index values from the current year. In the above example, this would look like the following:

$$Unadjusted\ 2018\ Index = \left(\frac{50}{125} * (-0.25) + \frac{50}{125} * 2.53 + \frac{25}{125} * 2.82\right) = 1.48 \quad (20)$$

Note that the index for each value-added measure is weighted according to the students associated with it. This teacher had 50 FTE students associated with the 2018 Geometry value-added measure, 50 FTE students associated with the 2018 Algebra II value-added measure, and 25 FTE students associated with the 2018 Algebra I value-added measure. The total number of FTE students totals 50 + 50 + 25, or 125. The index for 2018 Geometry (-0.25) is thus weighted proportionately at 50/125, the index for 2018 Algebra II (2.53) is also weighted at 50/125, and the index for 2018 Algebra I (2.82) is weighted at 25/125. In equation (20) above and all other evaluation composite calculations, the unrounded index values are used (meaning, the value-added measure divided by its standard error rather than the rounded value reported in Table 7).

Since each of the individual index values have a standard error of 1, there needs to be an additional correction to recalculate the overall average index to make it have a standard error of 1 or so that it is standardized like the original index values. This standard error of an average index can be found using the following formula:

$$SE\ for\ 2018\ Index = \sqrt{\left(\frac{50}{125}\right)^2 + \left(\frac{50}{125}\right)^2 + \left(\frac{25}{125}\right)^2} = 0.6 \quad (21)$$

To calculate the new index, the average of the index values would be divided by the new standard error of the average index.

$$Final\ 2018\ Index = \frac{1.48}{0.6} = 2.46 \quad (22)$$

Notice how the index value of the composite is larger than the average index. This is because there is more information and evidence about students' growth when all the individual measures are combined. The additional evidence provides a greater level of certainty that this teacher's students are demonstrating above average growth across the subjects and grades in the current year.

[5] See, for example, Dennis D. Wackerly, William Mendenhall III, and Richard L. Scheaffer, "Chapter 7" in *Mathematical Statistics with Applications, Sixth Edition* (Pacific Grove, CA: Duxbury Thomson Learning, Inc., 2002).

6.1.3 Calculation of the multi-year evaluation composite

The multi-year evaluation composite includes up to three years of value-added measures. This could include the single-year composite index based on 2017-18 reporting (weighted at 10% of the 35% allocated to growth in the teacher evaluation), 2016-17 reporting (weighted at 10% of the 35%), and 2015-16 (weighted at 15% of the 35%). If a teacher does not have value-added measures from 2016-17, then the 2015-16 data would be weighted at 25% instead of 15%. The sample below assumes that the teacher has all three years of data. To calculate this composite, the single-year composite calculated in the previous section would be weighted at 10%, and similar steps would be taken with the measures from the two prior years.

The first step is to calculate 2017 and 2016 adjusted index values similar to what is done above for 2018.

$$Unadjusted\ 2017\ Index = \left(\frac{50}{100} * 0.36 + \frac{50}{100} * 2.81\right) = 1.58 \tag{23}$$

$$SE\ for\ 2017\ Index = \sqrt{\left(\frac{50}{100}\right)^2 + \left(\frac{50}{100}\right)^2} = 0.71 \tag{24}$$

$$Final\ 2017\ Index = \frac{1.58}{0.71} = 2.24 \tag{25}$$

Again, within the 2017 index, the indices for each value-added measure are weighted according to the students associated with it. This teacher had 50 FTE students associated with the 2017 Algebra I value-added measure and 50 FTE students associated with the 2017 Algebra II value-added measure. The total number of FTE students totals 50 + 50, or 100.

$$Unadjusted\ 2016\ Index = \left(\frac{25}{125} * 2.17 + \frac{100}{125} * 2.33\right) = 2.30 \tag{26}$$

$$SE\ for\ 2016\ Index = \sqrt{\left(\frac{25}{125}\right)^2 + \left(\frac{100}{125}\right)^2} = 0.82 \tag{27}$$

$$Final\ 2016\ Index = \frac{2.30}{0.82} = 2.79 \tag{28}$$

Similarly, for the 2016 index, the index for 2016 Biology I (2.17) is thus weighted proportionately at 25/125, and the index for 2016 Algebra II (2.33) is weighted at 100/125.

Before combining the individual years into a multi-year index, each year's index is adjusted as in the single year composite. The standard error for the 2017 and 2016 unadjusted index value is 0.71 and 0.82, respectively. These are calculated in the same way as was done for the 2018 single year composite.

The next step is to calculate a multi-year index that combines the 2018, 2017, and 2016 indices according to their specified weights. This index is "unadjusted" and is not considered final until it is divided by its standard error.

$$Unadjusted\ Multi-year\ Index = \left(\frac{10}{35} * 2.46 + \frac{10}{35} * 2.24 + \frac{15}{35} * 2.79\right) = 2.54 \quad (29)$$

The standard error can again be calculated using the following formula, which accounts for the different weights of each year's index value in the overall multi-year index.

$$SE\ for\ Multi-year\ Index = \sqrt{\left(\frac{10}{35}\right)^2 + \left(\frac{10}{35}\right)^2 + \left(\frac{15}{35}\right)^2} = 0.59 \quad (30)$$

The new index value for the 2018, 2017 and 2016 would be as follows (using non-rounded numbers):

$$Final\ Multi-year\ Index = \frac{2.54}{0.59} = 4.31 \quad (31)$$

6.1.4 Calculation of the multi-year evaluation composite without 2015-16 data

The multi-year evaluation composite without 2015-16 data is similar to what was shown in the previous section. The single-year composite index based on the 2017-18 reporting is weighted at 10%, and another single-year composite based on 2016-17 reporting is also weighted at 10% for a total of 20% rather than 35%. Both single-year composites have been calculated in previous steps and can be used again here:

$$Unadjusted\ Multi-year\ Index = \left(\frac{10}{20} * 2.46 + \frac{10}{20} * 2.24\right) = 2.35 \quad (32)$$

The standard error can again be calculated using the following formula.

$$SE\ for\ Mult-year\ Index = \sqrt{\left(\frac{10}{20}\right)^2 + \left(\frac{10}{20}\right)^2} = 0.71 \quad (33)$$

The final index value for the multi-year composite without 2015-16 data would be as follows (using non-rounded numbers):

$$Final\ Multi-year\ Index = \frac{2.35}{0.71} = 3.33 \quad (34)$$

6.2 District and school evaluation composites

Districts and schools also receive evaluation composites. The TDOE policies for these composites are outlined below:

- District and school evaluation composites are single-year measures based entirely on the current year's reporting.
- District and school evaluation composites weigh the value-added measures that are included in the composite according to the number of students associated with each value-added measure.
- There are six types of evaluation composites: Overall, Numeracy, Literacy, a combined Numeracy and Literacy, Science, and Social Studies. These six types can be created using

different combinations of test data, and all options are listed in Section 6.2.1. Where applicable, the grades associated with each subject are included in parentheses.

6.2.1 Sample calculation of district/school evaluation composite

Like section 6.1.4, this section presents how school-level composites are calculated, and the decisions for schools share the same statistical approaches and policy decisions as those for teachers.

The key steps for determining a school's composite index are as follows:

1. Calculate MRM-based composite *gain*, *standard error,* and *index* across subjects and grades.
2. Calculate URM-based composite *index* across subjects.
3. Calculate *composite index* using both the MRM- and URM-based composite indices.

The following sections illustrate this process using value-added measures from a sample middle school, which are provided below:

Table 8: Sample School Value-Added Information

Year	Subject	Grade	Value-Added Gain	Standard Error	Number of Students
2018	Math	6	3.30	0.70	44
2018	ELA	6	-1.10	1.00	46
2018	Math	7	2.00	0.50	50
2018	ELA	7	2.40	1.10	50
2018	Math	8	-0.30	0.60	40
2018	ELA	8	3.80	0.70	50
2018	Algebra I	N/A	-11.50	6.20	35

6.2.2 Calculate MRM-based composite gain across subjects

As in the MRM-based composite gain for teachers, when the value-added estimates are in the same scale (Normal Curve Equivalents), the school composite gain across the six subject/grades is a weighted average based on the number of students in each subject and grade. For the school, the total number of students affiliated with MRM value-added measures is 44 + 46 + 50 + 50 + 40 + 50, or 280. The math grade 6 value-added measure would be weighted at 44/280, the ELA grade 6 value-added measure would be weighted at 46/280, and so on. More specifically, the composite gain is calculated using the following formula:

$$\begin{aligned} Comp\ Gain &= \tfrac{44}{280}Math_6 + \tfrac{46}{280}ELA_6 + \tfrac{50}{280}Math_7 + \tfrac{50}{280}ELA_7 + \tfrac{40}{280}Math_8 + \tfrac{50}{280}ELA_8 \\ &= \left(\tfrac{44}{280}\right)(3.30) + \left(\tfrac{46}{280}\right)(-1.10) + \left(\tfrac{50}{280}\right)(2.00) + \left(\tfrac{50}{280}\right)(2.40) + \left(\tfrac{40}{280}\right)(-0.30) + \\ &\quad \left(\tfrac{50}{280}\right)(3.80) = 1.76 \end{aligned} \tag{35}$$

6.2.3 Calculate MRM-based standard error across subjects

6.2.3.1 *Technical background on standard errors*

The standard error of the MRM school composite value-added gain cannot be calculated using the assumption that the gains making up the composite are independent. This is because many of the same students are likely represented in different value-added gains, such as grade 8 math in 2018 and grade 8 ELA in 2018. The statistical approach, outlined in Section 3.1.3 (with references), is quite sophisticated and will take into account the correlations between pairs of value-added gains as shown in equation (36) below and using equation (6) for schools and equation (10) for teachers.[6] The composites are indeed linear combinations of the fixed effects of the models and can be estimated as described in Section 3.1.3. The magnitude of each correlation depends on the extent to which the same students are in both estimates for any two subject/grade/year estimates.

6.2.3.2 *Illustration of MRM-based standard error for sample school*

As a reminder, the use of the word "error" does not indicate a mistake. Rather, value-added models produce *estimates*. The value-added gains in the above tables are estimates, based on student test score data, of the school's true value-added effectiveness. In statistical terminology a "standard error" is a measure of the uncertainty in the estimate, providing a means to determine whether an estimate is decidedly above or below the growth expectation. Standard errors can, and should, also be provided for the composite gains that have been calculated, as shown above, from a teacher's value-added gain estimate.

Statistical formulas are often more conveniently expressed as variances, and this is the square of the standard error. Standard errors of composites can be calculated using variations of the general formula shown below. To maintain the generality of the formula, the individual estimates in the formula (think of them as value-added-gains) are simply called X, Y, and Z. If there were more than or fewer than three estimates, the formula would change accordingly. As MRM composites use proportional weighting according to the number of students linked to each value-added gain, each estimate is multiplied by a different weight: a, b, or c.

$$\begin{aligned} Var(aX + bY + cZ) = a^2Var(X) + b^2Var(Y) + c^2Var(Z) \\ +2ab\ Cov(X,Y) + 2ac\ Cov(X,Z) + 2bc\ Cov(Y,Z) \end{aligned} \quad (36)$$

Covariance, denoted by Cov, is a measure of the relationship between two variables. It is a function of a more familiar measure of relationship, the correlation coefficient. Specifically, the term $Cov(X,Y)$ is calculated as follows:

$$Cov(X,Y) = Correlation(X,Y)\sqrt{Var(X)}\sqrt{Var(Y)} \quad (37)$$

The value of the correlation ranges from -1 to +1, and these values have the following meanings:

- A value of zero indicates no relationship.
- A positive value indicates a positive relationship, or Y tends to be larger when X is larger.
- A negative value indicates a negative relationship, or Y tends to be smaller when X is larger.

[6] For more details on the statistical approach to derive the standard errors, see, for example, Ramon C. Littell, George A. Milliken, Walter W. Stroup, Russell D. Wolfinger, and Oliver Schabenberger, *SAS for Mixed Models, Second Edition* (Cary, NC: SAS Institute Inc., 2006). Another example: Charles E. McCulloch, Shayle R. Searle, and John M. Neuhaus, *Generalized, Linear, and Mixed Models* (Hoboken, NJ: Wiley, 2008).

Two variables that are unrelated have a correlation and covariance of zero. Such variables are said to be statistically independent. If the X and Y values have a positive relationship, then the covariance will also be positive. As a general rule, two value-added gain estimates are statistically independent if they are based on completely different sets of students.

For our sample school's composite gain, the relationship will generally be positive, and this means that the MRM-based composite standard error is larger than it would be assuming independence. Using the student weightings and standard errors reported in Table 8 and assuming total independence, the standard error would then be as follows:

$$MRM\ Comp\ SE = \sqrt{\begin{matrix}\left(\frac{44}{280}\right)^2 (SE\ Math_6)^2 + \left(\frac{46}{280}\right)^2 (SE\ Read_6)^2 + \left(\frac{50}{280}\right)^2 (SE\ Math_7)^2 \\ + \left(\frac{50}{280}\right)^2 (SE\ Read_7)^2 + \left(\frac{40}{280}\right)^2 (SE\ Math_8)^2 + \left(\frac{50}{280}\right)^2 (SE\ Read_8)^2\end{matrix}}$$

$$= \sqrt{\begin{matrix}\left(\frac{44}{280}\right)^2 (0.70)^2 + \left(\frac{46}{280}\right)^2 (1.00)^2 + \left(\frac{50}{280}\right)^2 (0.50)^2 \\ + \left(\frac{50}{280}\right)^2 (1.10)^2 + \left(\frac{40}{280}\right)^2 (0.60)^2 + \left(\frac{50}{280}\right)^2 (0.70)^2\end{matrix}} = 0.33 \quad (38)$$

At the other extreme, if the correlation between each pair of value-added gains had its maximum value of +1, the standard error would be larger.

The actual standard error will likely be above the value of 0.33 due to students being in both math and ELA in the school with the specific value depending on the values of the correlations between pairs of value-added gains. Correlations of gains across years might be positive or slightly negative, as the same student's score can be used in multiple gains. The magnitude of each correlation depends on the extent to which the same students are in both estimates for any two subject/grade/year estimates.

For the sake of simplicity, let us assume the actual standard error was 0.40 for the school composite in this example.

6.2.4 Calculate MRM-based composite index across subjects

The next step is to calculate the MRM-based school composite index, which is the school composite value-added gain divided by its standard error. The MRM-based composite index for this school would be calculated as follows:

$$MRM\ Comp\ Index = \frac{MRM\ Comp\ Gain}{MRM\ Comp\ SE} = \frac{1.76}{0.40} = 4.40 \quad (39)$$

While some of the values in the example were rounded for display purposes, the actual rounding or truncating only occurs after all of measures have been combined, as described in Section 5.3.

6.2.5 Calculate URM-based index across subjects

For our sample school, there is only one available URM value-added measure. This means that the reported value-added index for that subject will be the same that is calculated for the URM-based composite index.

$$URM\ Comp\ Index = \frac{Alg\ I\ VA\ Measure}{Alg\ I\ SE} = \frac{-11.50}{6.20} = -1.85 \quad (40)$$

However, should a school or district have more than one value-added measure based on the URM, then the composite index would be calculated by first calculating index values for each subject and then combining those weighting by the number of students. The standard error of this combined index must assume independence since the URM measures are done in separate models for each year and subject

6.2.6 Calculate the combined MRM and URM composite index across subjects

The two composite indices from the MRM and URM are then weighted according to the number of students within each model to determine the combined composite index. Our sample school has 315 students, of which 280 are in the MRM and 35 in the URM. The combined composite index would be calculated as follows using these weightings, the MRM-based composite index across subjects, and the URM-based index across subjects:

$$Unadjusted\ Combined\ Comp\ Index = \left(\frac{280}{315}\right)4.40 + \left(\frac{35}{315}\right)(-1.85) = 3.71 \quad (41)$$

This combined index is not an actual index itself until it is adjusted to accommodate for the fact that it is based on multiple pieces of evidence together. An index, by definition, has a standard error of 1, but this unadjusted value (3.71) does not have a standard error of 1. The next step is to calculate the new standard error and divide the combined composite index found above by it. This new, adjusted composite index will be the final index with a standard error of 1. The standard error can be found given the standard formula above and the fact that each index has a standard error of 1. Independence is assumed since these are done outside of the models. In this example, the standard error would be as follows:

$$Final\ Combined\ Comp\ SE = \sqrt{\left(\frac{280}{315}\right)^2 (1)^2 + \left(\frac{35}{315}\right)^2 (1)^2} = 0.90 \quad (42)$$

Therefore, the final combined composite index value is 3.71 divided by 0.90, or 4.14. This is the value that determines the school evaluation composite. Different types of evaluation composites use the value-added measures from different tests, but the overall process is the same.

6.2.7 Types of evaluation composites

6.2.7.1 *TCAP (Grades 4-8)/EOC/Early Grades (Grade 3)*

Composite Type	Subjects
Overall	Algebra I, Algebra II, Biology I, English I, English II, English III, Chemistry, Geometry, Integrated Math I, Integrated Math II, Integrated Math III, TCAP Math/English Language Arts (3 and 4–8), TCAP Science/Social Studies (5–8), and U.S. History
Literacy	English I, English II, English III, and TCAP English Language Arts (3 and 4–8)
Numeracy	Algebra I, Algebra II, Geometry, Integrated Math I, Integrated Math II, Integrated Math III, and TCAP Math (3 and 4–8)
Literacy and Numeracy	Algebra I, Algebra II, English I, English II, English III, Geometry, Integrated Math I, Integrated Math II, Integrated Math III, and TCAP Math/English Language Arts (3 and 4–8)

Composite Type	Subjects
Science	Biology I, Chemistry, and TCAP Science (5–8)
Social Studies	TCAP Social Studies (5–8) and U.S. History

6.2.7.2 *TCAP (Grades 4-8)/EOC*

Composite Type	Subjects
Overall	Algebra I, Algebra II, Biology I, English I, English II, English III, Chemistry, Geometry, Integrated Math I, Integrated Math II, Integrated Math III, TCAP Math/English Language Arts (4–8), TCAP Science/Social Studies (5–8), and U.S. History
Literacy	English I, English II, English III, and TCAP English Language Arts (4–8)
Numeracy	Algebra I, Algebra II, Geometry, Integrated Math I, Integrated Math II, Integrated Math III, and TCAP Math (4–8)
Literacy and Numeracy	Algebra I, Algebra II, English I, English II, English III, Geometry, Integrated Math I, Integrated Math II, Integrated Math III, and TCAP Math/English Language Arts (4–8)
Science	Biology I, Chemistry, and TCAP Science (5–8)
Social Studies	TCAP Social Studies (5–8) and U.S. History

6.2.7.3 *TCAP (Grades 4-8)*

Composite Type	Subjects
Overall	TCAP Math/English Language Arts (4–8) and TCAP Science/Social Studies (5–8)
Literacy	TCAP English Language Arts (4–8)
Numeracy	TCAP Math (4–8)
Literacy and Numeracy	TCAP Math/English Language Arts (4–8)
Science	TCAP Science (5–8)
Social Studies	TCAP Social Studies (5–8)

6.2.7.4 *EOC*

Composite Type	Subjects
Overall	Algebra I, Algebra II, Biology I, English I, English II, English III, Chemistry, Geometry, Integrated Math I, Integrated Math II, Integrated Math III
Literacy	English I, English II, English III
Numeracy	Algebra I, Algebra II, Geometry, Integrated Math I, Integrated Math II, Integrated Math III

Composite Type	Subjects
Literacy and Numeracy	Algebra I, Algebra II, English I, English II, English III, Geometry, Integrated Math I, Integrated Math II, Integrated Math III
Science	Biology I, Chemistry
Social Studies	U.S. History

6.2.7.5 *CTE Students (Based on EOC)*

Composite Type	Subjects
Overall	Algebra I, Algebra II, Biology I, English I, English II, English III, Chemistry, Geometry, Integrated Math I, Integrated Math II, Integrated Math III, and U.S. History
Literacy	English I, English II, and English III
Numeracy	Algebra I, Algebra II, Geometry, Integrated Math I, Integrated Math II, and Integrated Math III
Literacy and Numeracy	Algebra I, Algebra II, English I, English II, English III, Geometry, Integrated Math I, Integrated Math II, and Integrated Math III
Science	Biology I and Chemistry
Social Studies	U.S. History

6.2.7.6 *CTE Concentrators (Based on EOC)*

Composite Type	Subjects
Overall	Algebra I, Algebra II, Biology I, English I, English II, English III, Chemistry, Geometry, Integrated Math I, Integrated Math II, Integrated Math III, and U.S. History
Literacy	English I, English II, and English III
Numeracy	Algebra I, Algebra II, Geometry, Integrated Math I, Integrated Math II, and Integrated Math III
Literacy and Numeracy	Algebra I, Algebra II, English I, English II, English III, Geometry, Integrated Math I, Integrated Math II, and Integrated Math III
Science	Biology I and Chemistry
Social Studies	U.S. History

6.2.7.7 *Early Grades (Grade 3)*

Composite Type	Subjects
Overall	TCAP Math/English Language Arts (3)
Literacy	TCAP English Language Arts (3)

Composite Type	Subjects
Numeracy	TCAP Math (3)
Literacy and Numeracy	TCAP Math/English Language Arts (3)

6.2.8 District and school subgroup composites

As described in Sections 3.1.5.1 and3.2.4.2, Tennessee uses subgroup-level value-added measures in their federal accountability system. For the subgroups described in these sections, the MRM and URM growth measures are combined in the same way as the overall measure described in Section 6.2.1 through 6.2.6.

District measures are available for the following grades in math, English language arts and science:

- Grades 3–5 (for districts that administer grade 2)
- Grades 4–5 (for all districts)
- Grades 6–8
- Grades 9–12

School measures are available for the following grades in math, English language arts and science:

- All grades at the school (including grade 3 if applicable and meets grade 2 criteria)
- All grades in the school, not including grade 3

Depending on the eligible growth measures for the district and school, growth measures from the following assessments might be included:

Composite Type	Subjects
Math	TCAP Math in grades 3–8, Algebra I, Algebra II, Integrated Math I, Integrated Math II, Integrated Math III, Geometry
English language arts	TCAP English language arts in grades 3–8, English I, English II, English III
Science	TCAP Science in grades 5–8

7 TVAAS Projection Model

7.1 Available projections

In addition to providing value-added modeling, TVAAS provides a variety of additional services including projected scores for individual students on tests that they have not yet taken. These tests include all assessments that are used in value-added in the state of Tennessee. These projections can be used to predict a student's future success or lack thereof. As such, this projection information can be used as an early warning indicator to guide counseling and intervention to increase students' likelihood of future success.

The following projections are available to educators in Tennessee within the 2017-18 reporting:

- Math and English language arts in grades 3–8;
- Science and social studies in grades 5–8;
- EOC Algebra I, Algebra II, Biology I, English I, and English II;
- ACT Composite, English, math, reading and science.

Although the projection model was modified for the 2016-17 reporting due to the missing 2015-16 test data, the 2017-18 reporting will follow the more tradition model used in years prior to 2016-17. When making projections to grades 4 and 5 in 2017-18, only two prior test scores are required as predictors instead of the typical three due to the test scores available.

7.2 Modeling approach

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:

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i \quad (43)$$

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:

$$\hat{y}_i = \hat{\mu}_y \pm \hat{\alpha}_j + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots + \epsilon_i \quad (44)$$

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 might prefer to provide a list of feeder patterns from which it is possible to determine the most likely school 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 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 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 (or at least two available predictor scores for projections to grades 4 and 5). 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. For social studies, the projections will not provide probabilities to specific performance levels since those levels will not be available at the time of release. Rather, the initial projections will be based on the probability of obtaining a particular percentile.

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.

$$Prob(\hat{y}_i \geq b) = \Phi\left(\frac{\hat{y}_i - b}{SE(\hat{y}_i)}\right) \quad (45)$$

8 Data quality and pre-analytic data processing

This section provides an overview of the steps taken to ensure sufficient data quality and processing for reliable value-added analysis.

8.1 Data quality

Data are provided each year to EVAAS consisting of student test data and file formats. These data are checked each year to be incorporated into a longitudinal database that links students over time. Student test data and demographic data are checked for consistency year to year to assure that the appropriate data are assigned to each student. Student records are matched over time using all data provided by the state. Teacher records are matched over time using the TLN and teacher's name.

8.2 Checks of scaled score distributions

The statewide distribution of scale scores is examined each year to determine whether they are appropriate to use in a longitudinally linked analysis. Scales must meet the three requirements listed in Section 2.1 and described again below to be used in all types of analysis done within TVAAS. Stretch and reliability are checked every year using the statewide distribution of scale scores that is sent each year before the full test data is given.

8.2.1 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 measurable differences to be observed. Stretch can be determined by the percentage of students who score near the minimum or the maximum level for each assessment. If a large percentage of students scored at the maximum in one grade compared to the prior grade, then it might 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. Percentages for the Grade 2, TCAP and EOC Assessments are suitable for value-added analysis, meaning that the state tests have adequate stretch to measure value-added even in situations where the group of students are very high or low achieving.

8.2.2 Relevance

Relevance indicates whether the test has sufficient alignment with the state standards. The requirement that 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. This is how the state tests are designed and is monitored by the TDOE and their psychometricians.

8.2.3 Reliability

Reliability can be viewed in a few different ways for assessments. Psychometricians view reliability as the idea that student would receive similar scores if they took the assessment multiple times. This type of reliability is important for most any use of standardized assessments. Reliability also refers to the assessment's scales across years. This second type of reliability is very important if 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. (Tennessee historically used a base-year approach for value-added reports in TCAP grades 4–8 until the year 2014-15. The value-added model now uses an intra-year approach.) Both types of reliability are important when measuring growth.

8.3 Data quality business rules

The pre-analytic processing regarding student test scores is detailed below.

8.3.1 Missing grade levels

In Tennessee, the grade level used in the analyses and reporting is the tested grade, not the enrolled grade. If a grade level is missing on any grade-level tests (meaning K–8), then these records will be excluded from all analyses. The grade is required to include a student's score into the appropriate part of the models, and it would need to be known if the score was to be converted into an NCE.

8.3.2 Duplicate (same) scores

If a student has a duplicate score for a particular subject and tested grade in a given testing period in a given school, then the extra score will be excluded from the analysis and reporting.

8.3.3 Students with missing districts or schools for some scores but not others

If a student has a score with a missing district or school for a particular subject and grade in a given testing period, then the duplicate score that has a district and/or school will be included over the score that has the missing data. This rule applies individually to specific subject/grade/years.

8.3.4 Students with multiple (different) scores in the same testing administration

If a student has multiple scores in the same period for a particular subject and grade and the test scores are not the same, then those scores will be excluded from the analysis. If duplicate scores for a particular subject and tested grade in a given testing period are at different schools, then both of these scores will be excluded from the analysis.

8.3.5 Students with multiple grade levels in the same subject in the same year

A student should not have different tested grade levels in the same subject in the same year. If that is the case, then the student's records are checked to see whether the data for two separate students were inadvertently combined. If this is the case, then the student data are adjusted so that each unique student is associated with only the appropriate scores. If the scores appear to all be associated with a single unique student, then scores that appear inconsistent are excluded from the analysis.

8.3.6 Students with records that have unexpected grade level changes

If a student skips more than one grade level (e.g., moves from sixth last year to ninth this year) or is moved back by one grade or more (i.e. moves from fourth grade last year to third this year) in the same subject, then the student's records are examined to determine whether two separate students were inadvertently combined. If this is the case, then the student data is adjusted so that each unique student is associated with only the appropriate scores. If it is the same student, then these scores are removed from the analysis.

8.3.7 Students with records at multiple schools in the same test period

If a student is tested at two different schools in a given testing period, then the student's records are examined to determine whether two separate students were inadvertently combined. If this is the case, then the student data is adjusted so that each unique student is associated with only the appropriate scores. When students have valid scores at multiple schools in different subjects, all valid scores are used at the appropriate school.

8.3.8 Outliers

8.3.8.1 *Conceptual Explanation*

Student assessment scores are checked each year to determine whether any scores are "outliers" in context with all other scores in a reference group of scores from an individual student. This is one of the protections in place with TVAAS analyses and reporting. This is a conservative process by which scores are statistically examined to determine whether a score is considered an outlier. In other words, is the score "significantly different" from the other scores, as indicated by a statistical analysis that compares each score to the other scores? There are different business rules for the low outlier scores and the high outlier scores, and this approach is more conservative when removing a very high achieving score. In other words, a lower score would be considered an outlier before a higher score would be considered an outlier. Again, this is a protection with TVAAS.

8.3.8.2 *Technical Explanation*

Student assessment scores are checked each year to determine whether they are outliers in context with all other scores in a reference group of scores from the individual student. These reference scores are weighted differently depending on proximity in time to the score in question. Scores are checked for outliers using related subjects as the reference group. For example, when searching for outliers for math test scores, all math subjects (early grade, TCAP and EOC assessments) are examined simultaneously,

and any scores that appear inconsistent, given the other scores for the student, are flagged. Note that grade 2 scores are used to detect outliers for grade 3 scores.

Scores are flagged in a conservative way to avoid excluding any student scores that should not be excluded. Scores can be flagged as either high or low outliers. Once an outlier is discovered, that outlier will not be used in the analysis, but it will be displayed on the student testing history on TVAAS web application.

This process is part of a data quality procedure to ensure that no scores are used if they were in fact errors in the data, and the approach for flagging a student score as an outlier is fairly conservative.

Considerations included in outlier detection are:

- Is the score in the tails of the distribution of scores? Is the score very high or low achieving?
- Is the score "significantly different" from the other scores, as indicated by a statistical analysis that compares each score to the other scores?
- Is the score also "practically different" from the other scores? Statistical significance can sometimes be associated with numerical differences that are too small to be meaningful.
- Are there enough scores to make a meaningful decision?

To decide whether student scores are considered outliers, all student scores are first converted into a standardized normal z-score. Then, each individual score is compared to the weighted combination of all the reference scores described above. The difference of these two scores will provide a t-value of each comparison. This t-value provides information as to how many standard deviations away the score is from the weighted combination of all reference scores. Using this t-value, EVAAS can flag individual scores as outliers.

There are different business rules for the low outliers and the high outliers, and this approach is more conservative when removing a very high achieving score.

For low-end outliers, the rules are:

- The percentile of the score must be below 50.
- The t-value must be below -2.5 when determining the difference between the score in question and the weighted combination of reference scores (otherwise known as the comparison score). In other words, the score in question must be at least 2.5 standard deviations below the comparison score.
- The percentile of the comparison score must be above a certain value. This value depends on the position of the individual score in question but will range from 10 to 90 with the ranges of the individual percentile score.

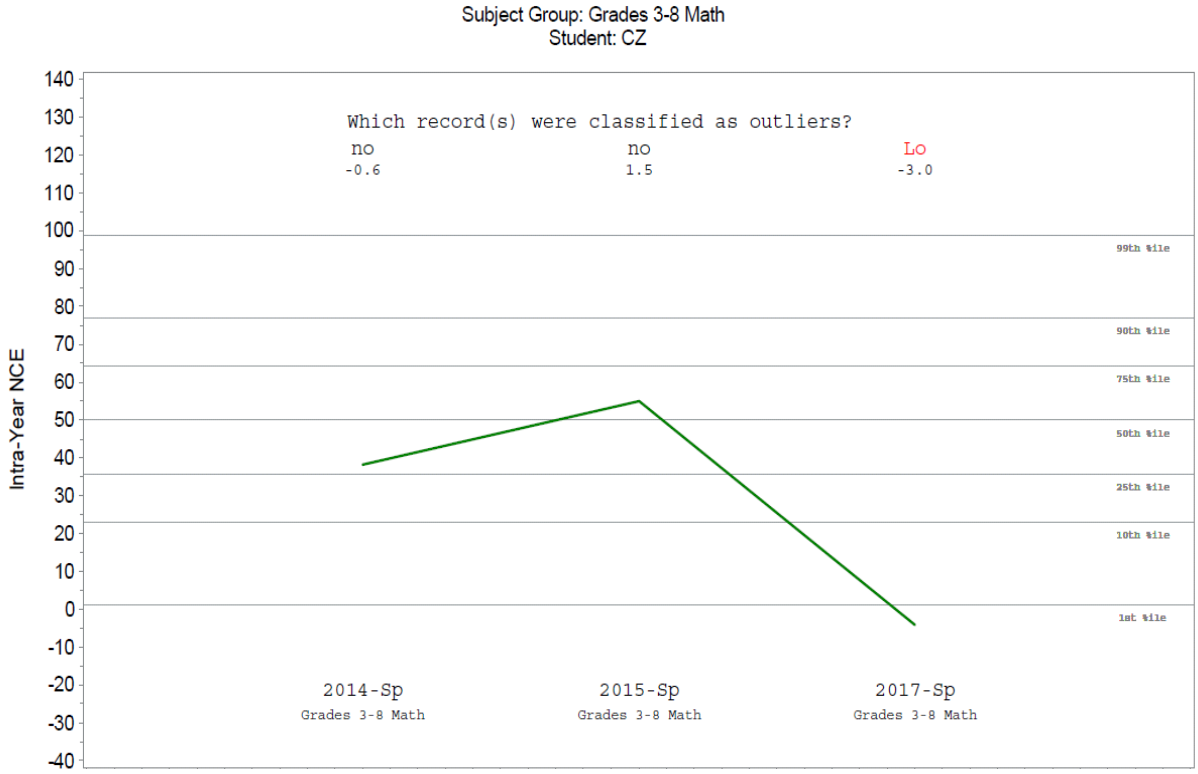
For high-end outliers, the rules are:

- The percentile of the score must be above 50.
- The t-value must be above 4.5 when determining the difference between the score in question and the reference group of scores. In other words, the score in question must be at least 4.5 standard deviations above the comparison score.
- The percentile of the comparison score must be below a certain value. This value depends on the position of the individual score in question but will range from 20 to 50 with the ranges of

the individual percentile score. There must be at least three reference scores used to make the comparison score.

The figure below provides a visual example of this process. A student's annual scores for math are plotted on the graph. The left y-axis reports the student scores in intra-year NCE units while the right y-axis reports the student scores in percentiles. It is clear the student's 2017 math score is lower than the student's previous scores, and, using the process outlined above in conjunction with all of the student's scores from other subjects, the 2017 math score is determined to be an outlier. It is marked as "Lo" in red at the top. The numbers at the top represent the t-values discussed above.

Figure 3: Outlier detection example



If there are any additional questions regarding the information in this document, click here to go to the TVAAS Contact Us page for additional resources.