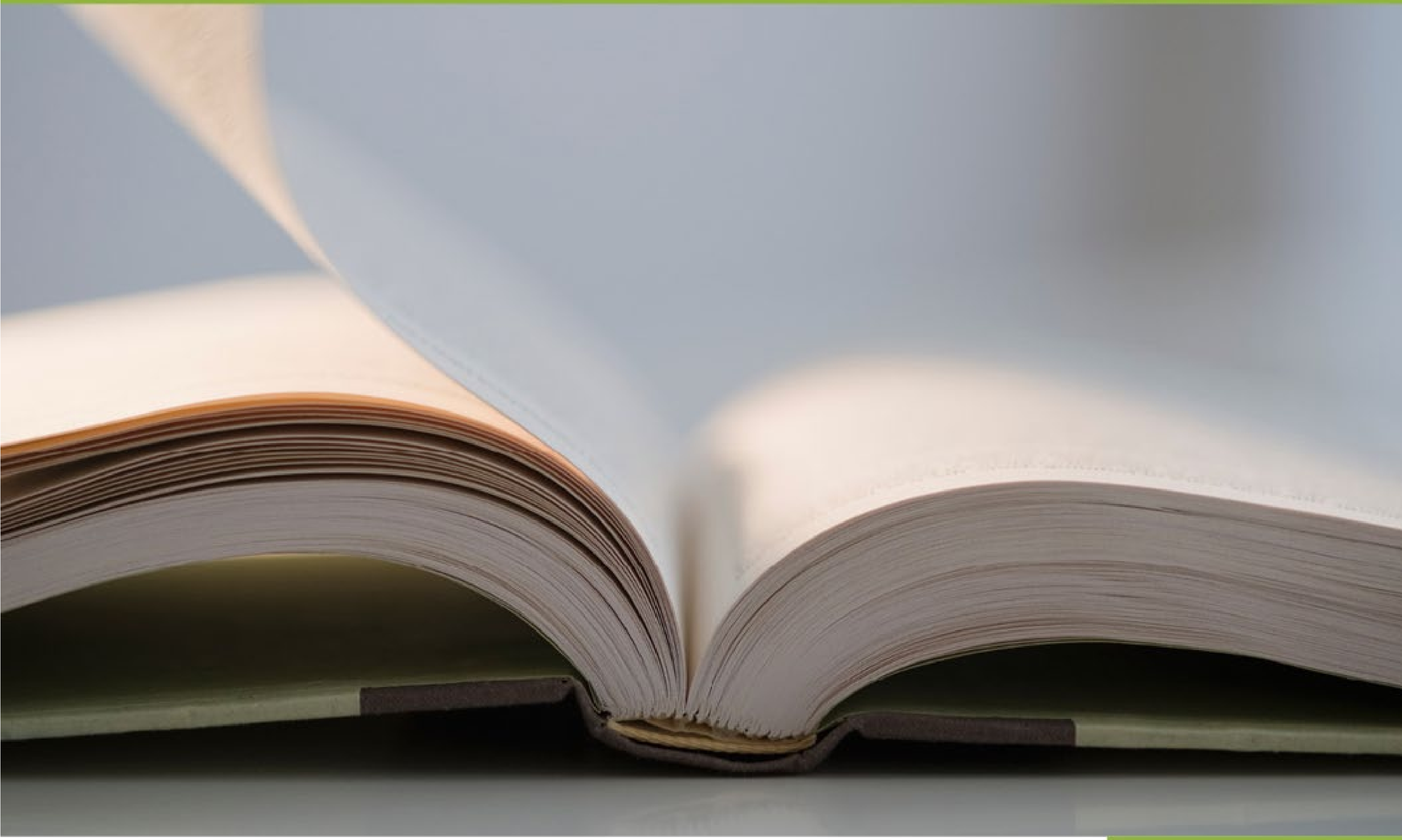
SAS® EVAAS®

Statistical Models and Business Rules of OH EVAAS Analyses

Prepared for Ohio Department of Education



THE POWER TO KNOW.

Contents

1 Introduction to value-added reporting in Ohio

The term "value-added" refers to a statistical analysis used to measure the impact of districts, schools, and teachers on the academic progress rates of groups of students from year to year. Conceptually and as a simple explanation, a value-added "score" is calculated in the following manner:

- Growth = current achievement/current results compared to prior achievement/prior results with achievement being measured by a quality assessment such as Ohio's state tests (OSTs).

While the concept of growth is easy to understand, the implementation of a statistical model of growth is more complex. There are several 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 the 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 guide you through value-added modeling based on the statistical approaches, policies, and practices selected by the State of Ohio and currently implemented by EVAAS. This document describes the input data, modeling, and business rules for district, school, and teacher value-added reporting in Ohio.

The State of Ohio and the EVAAS team have provided value-added reporting since 2002. The initial collaboration was through Project SOAR, a 42-district pilot. By 2006, district and school value-added reporting was available statewide, and in 2011, teacher value-added reports also became available for parts of the state. The first year of statewide implementation for teacher value-added reporting that included all teachers with students taking the state assessments in grades 4–8 was 2013.

2 Input data used in the Ohio value-added model

This section provides details regarding the input data used in the Ohio value-added model, such as the requirements for verifying appropriateness in value-added analysis as well as the student, teacher, principal, and/or 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 on each of these requirements are provided in Section 8, Data quality and pre-analytic data processing.)

- **There is sufficient stretch in the scales** to ensure 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 highly related to the academic standards** so that it is possible to measure progress with the assessment in that subject/grade/year.
- **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 met by Ohio's standardized assessments and all vendor assessments that EVAAS receives from Ohio.

The current value-added implementation includes statewide assessments such as the OSTs, which measure Ohio's standards, as well as extended testing for some districts (vendor assessments that measure subjects and grades outside the state testing scope). There is potential to provide value-added reporting based on norm-referenced, college and career readiness, and end-of-course assessments.

2.1.2 Transitioning to future assessments

Ohio implemented new assessments in the 2014-15 school year and again in the 2015-16 school year. There were no test changes in the 2016-17 or 2017-18 school years. 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 providing value-added and growth reporting to 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 some relationship between the new test and old test. For example, a new grade 6 math exam should be correlated to previous math scores in grades 4 and 5 and to a lesser extent other grades and subjects such as reading 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 Ohio

Ohio's state tests (OSTs) are administered in the spring semester except for reading in grade 3, which is given in the fall and/or spring semesters. In grade 3, the higher of the two scores for each student are used in the value-added reporting, which is consistent with the accountability rules in Ohio. The ACT or SAT assessment is administered to all students across the state in the spring of grade 11.

2.2.1 Tests given in consecutive grades for the same subject

EVAAS receives tests that are given in consecutive grades for the same subject, which currently include:

- OST mathematics in grades 3–8
- OST reading in grades 3–8

2.2.2 Tests given in non-consecutive grades for the same subject

EVAAS receives tests that are given in non-consecutive grades for the same subject, which currently include:

- OST Algebra I, Mathematics I and II, and Geometry
- OST English Language Arts (ELA) I and II
- OST science in grades 5 and 8
- OST Biology
- OST American History and American Government
- ACT Math, English, and Reading
- SAT Evidence-Based Reading and Writing, and Math

2.2.3 Student identification information

Ohio's state law prohibits ODE from maintaining student names; therefore, the data ODE sends to EVAAS contains only the state student ID (SSID) for each student and no name information. IBM contracts with the State of Ohio to maintain the crosswalk with student names and IDs, so IBM securely transfers student names to Battelle for Kids (BFK) and The Management Council of the Ohio Education Computer Network (MCOECN). Those student names are matched using SSID and forwarded to EVAAS. These data are populated in the secure EVAAS website and then accessed by Local Education Agencies (LEAs) for further analysis and improvement purposes. The file from IBM contains the following:

- Student last name
- Student first name
- Student date of birth
- State student ID (SSID)

2.2.4 Assessment information provided

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

- Scale score
- Performance level
- Test taken
- Tested grade

- Accountable district IRN
- Accountable org IRN
- Testing district IRN
- Testing org IRN
- Reporting district IRN
- Reporting org IRN

Some of this information, such as performance levels, is not relevant to the ACT or SAT tests.

2.3 Student-level information

Student-level information is used in creating reports displayed in the EVAAS web application, so educators can analyze the data to inform practice and assist all students with academic progress. This information is also used for accountability categories that are reported to the public. EVAAS receives this information in the form of various socioeconomic, demographic, and programmatic identifiers in the student data system. In some cases, these identifiers are used to create categories for the accountability system. Currently, these categories are:

- Gifted – Reading
- Gifted – Math
- Gifted – Science
- Gifted – Superior Cognitive
- Migrant
- Limited English Proficient
- Economically Disadvantaged
- Students with Disabilities
- Gender
- Race
    - American Indian
    - Asian/Pacific Islander
        - This includes Asian and Hawaiian/Other Pacific Islander
    - Black
    - Hispanic
    - White
    - Multi-Racial

More information can be found in Ohio's Education Management Information System (EMIS) Manual about each of these identifiers and how they are defined by ODE at: http://education.ohio.gov/Topics/Data/EMIS/EMIS-Documentation/Current-EMIS-Manual.

2.4 Teacher-level 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-level value-added scores are calculated, teachers in Ohio are given the opportunity to complete roster verification to verify linkages between themselves and their students during the year. Roster verification by the individual teachers is an important part of a valid system. Roster verification enables teachers to confirm their class rosters for students they taught for a particular subject, grade, and year. These linkages, or records of teacher responsibility for specific students in specific subjects and grades, are verified by administrators as an additional check. The roster verification process also captures different teaching scenarios where multiple teachers can share instruction. Verification therefore increases the reliability and accuracy of teacher-level analyses.

For the purposes of Ohio's teacher-level value-added reports, EVAAS receives teacher identification data and student-teacher linkages from both BFK and MCOECN. The roster verification process provides data about the percentage of instructional responsibility of each teacher that may be attributed to a student.

The information contained in the student-teacher linkage files includes the following:

- District IRN
- District name
- School IRN
- School name
- Teacher level identification
    - Teacher name
    - Teacher state ID
- Student linking information, including SSID
- Subjects
- Percentage claimed by teacher

Whenever districts do not participate in roster verification, the teacher-student linkage reported and verified through EMIS is sent and used by EVAAS.

2.5 Principal-level information

EVAAS receives two data files from ODE on individual principals and assistant principals linking each of them to their schools. One provides a listing of principals and assistant principals in every school, their employment information, and their position start and end dates for that position, reported into EMIS by districts/community schools/JVSDs/ESCs. The other data file provides a listing of principals and assistant principals in every school, their employment information, and the school year in which they are reported in that position by districts/community schools/JVSDs/ESCs.

The term "principal" here refers to both assistant principals and principals. They are equivalent for the purposes of the calculations done later that are detailed in the composites section.

2.6 Data files by source

Table 1: Data Files Received by EVAAS

Source	Data
Ohio Department of Education	Student-level assessment data
Battelle for Kids	Teacher-student linkages
Management Council	Teacher-student linkages

IBM	Student names and SSIDs

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 prior achievement/prior results with achievement being measured by a quality assessment such as the OSTs.

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 Ohio, EVAAS provides two main categories of value-added models, each comprised of district, school and teacher level reports.

- **Multivariate Response Model (MRM)** is used for tests given in consecutive grades, like OST math and reading assessments in grades 3–8.
- **Univariate Response Model (URM)** is used when a test is given in non-consecutive grades, such as OST science assessments in grades 5 and 8 or any End-of-Course tests.

Both models offer the following advantages:

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

Each model is described in greater detail below.

As a result of using multiple years of test scores for each student and including students even if they have missing test scores, 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 regards 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 Ohio does not make any direct adjustments for students' socioeconomic/demographic characteristics. **Through this approach, Ohio avoids the problem of building a system that creates differential expectations for groups of students based on their backgrounds.**

The value-added reporting in Ohio is available at the district, school, and teacher level.

3.1 Multivariate Response Model reporting for tests in consecutive grades

EVAAS provides three separate analyses using the Multivariate Response Model (MRM), 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 statistical models are 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 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 4.1.)

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

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

Because of these advantages, the MRM is considered to be one of the most statistically robust and reliable approaches. The references below include recent 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 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?

3.1.1 MRM at the conceptual level

The example data in Figure 1 might help users understand how the MRM works. Assume that 10 students are given a test in two different years, with the results shown in Figure 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 are no missing data, these two simple methods provide the same answer (5.8 on the left in Figure 1). When there are missing data, though, each method provides a different result (6.9 vs. 4.6 on the right in Figure 1). A more sophisticated model is needed to address missing data.

Figure 1: Scores without missing data, and scores with missing data

Student	Previous Score	Current Score	Gain
1	51.9	74.8	22.9
2	37.9	46.5	8.6
3	55.9	61.3	5.4
4	52.7	47.0	-5.7
5	53.6	50.4	-3.2
6	23.0	35.9	12.9
7	78.6	77.8	-0.8
8	61.2	64.7	3.5
9	47.3	40.6	-6.7
10	37.8	58.9	21.1
Mean	50.0	55.8	5.8
	Difference	5.8	

Student	Previous Score	Current Score	Gain
1	51.9	74.8	22.9
2		46.5	
3	55.9	61.3	5.4
4		47.0	
5	53.6	50.4	-3.2
6	23.0	35.9	12.9
7	78.6	77.8	-0.8
8	61.2	64.7	3.5
9	47.3	40.6	-6.7
10	37.8	58.9	21.1
Mean	51.2	55.8	6.9
	Difference	4.6	

The MRM uses the correlation between current and previous scores in the non-missing data to estimate means for all previous and current scores as if there were no missing data. It does this without explicitly imputing values for the missing scores. The difference between these two estimated means is an estimate of the average gain for this group of students. In this example, the estimated difference is 5.8. Even in a small example such as this, the estimated difference is much closer to the difference with no missing data than either measure obtained by the mean of the differences (6.9) or the difference of the means (4.6). This method of estimation has been shown, on average, to outperform both of the simple methods.[1] This small example only considered two grades and one subject. Larger data sets, such as

[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. Available online at https://pvaas.sas.com/support/EVAAS-AdvantagesOfAMultivariateLongitudinalApproach.pdf.

those used in actual EVAAS analyses for Ohio, provide better correlation estimates by having more student data and more 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 concepts of the school and district models. The teacher model is slightly more complex, and all models are explained in more detail below (in Section 3.1.3). The first step in the MRM is to define the scores that will be used in the model.

3.1.2 Normal curve equivalents

3.1.2.1 Why EVAAS uses normal curve equivalents in MRM

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

NCEs are on a familiar scale because they are scaled to look like percentiles. However, NCEs have a critical advantage for measuring growth: they are on an equal-interval scale. This means that for NCEs, unlike percentile ranks, the distance between 50 and 60 is the same as the distance between 80 and 90. NCEs are constructed to be equivalent to percentile ranks at 1, 50, and 99, with the mean being 50 and the standard deviation being 21.063 by definition. Although percentile ranks are usually truncated above 99 and below 1, NCEs are allowed to range above 100 and below 0 to preserve their equal-interval property and to avoid truncating the test scale. See Figure 2 below for an illustration of the distribution of test scores, percentiles, and NCEs.

Figure 2: Percentile and NCE distributions

In a typical year in Ohio, the average maximum NCE is approximately 125. For display purposes in the EVAAS web application, NCEs are shown as integers from 1-99. Truncating would create an artificial ceiling or floor which may bias the results of the value-added measure for certain types of students forcing the gain to be close to zero or even negative.

The NCEs used in EVAAS analyses are based on a reference distribution of test scores in Ohio. The *reference distribution* is the distribution of scores on a state-mandated test for all students in either a given year (the base year approach) or in each year (intra-year approach). The base year (set in 2010) was previously used in the Ohio MRM analysis, and the intra-year approach was used for the first time in the 2014-15 reporting, as it can accommodate a change in testing regime when the old test and new test are not on the same scale.

By definition, the mean (or average) NCE score for the reference distribution is 50 for each grade and subject. "Growth" is the difference in NCEs from one year/grade to the next in the same subject. The growth standard, which represents a "normal" year's growth, is defined by a value of zero. More specifically, it maintains the same position in the reference distribution from one year/grade to the next. **It is important to reiterate that a gain of zero on the NCE scale does not indicate "no growth." Rather, it indicates that a group of students in a district, school, or classroom has maintained the same**

position in the state distribution from one grade to the next. The expectation of growth can be set differently by using a reference distribution to create NCEs or by using each individual year to create NCEs. For more on Growth Expectation, see Section 4.

3.1.2.2 How EVAAS uses normal curve equivalents in MRM

There are multiple ways of creating NCEs. EVAAS uses a method that does not assume the underlying scale is normal since experience has shown that some testing scales are not normally distributed, and this will ensure an equal interval scale. Table 2 provides an example of the way that EVAAS converts scale scores to NCEs.

The first five columns of Table 2 show an example of a tabulated distribution of test scores from Ohio data. The tabulation shows for each possible test score in a particular subject, grade, and year, how many students made that score ("Frequency") and what percent ("Percent") that frequency was out of the entire student population. (In Table 2, 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 2. If a particular score has a percentile rank of 48, this is interpreted to mean that 48% of students in the population had a lower score and 52% had a higher score. In practice, a non-zero percentage of students will receive each specific score; for example, 3.4% of students received a score of 425 in Table 2. The usual convention is to consider half of that 3.4% to be "below" and half "above." Adding 1.7% (half of 3.4%) to the 43.5% who scored below the score of 425 produces the percentile rank of 45.2 in Table 2.

Table 2: Converting tabulated test scores to NCE values

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

NCEs are obtained from the percentile ranks using the normal distribution. Using a table of the standard normal distribution (found in many textbooks[2]) 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

[2] See, for example, the inside front cover of William Mendenhall, Richard L. Scheaffer, and Dennis D. Wackerly, *Mathematical Statistics with Applications* (Boston: Duxbury Press, 1986).

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:

$$y = X\beta + Zv + \epsilon \tag{1}$$

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

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

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

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

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

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

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

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

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:[3]

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

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

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

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

1. Equation (5) below provides the best linear unbiased estimator (BLUE) of the set of estimable linear function, $K^T\beta$, of the fixed effects. The second equation (6) below represents the variance

[3] 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.

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.

3.1.3.1 District and school level

The district and school MRMs do not contain random effects; consequently, the Zv term drops out in the linear mixed model. 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" covariance matrix R_0 that contains a row and column for each subject and grade. Covariances among subjects and grades are assumed to be the same for all years (technically, all cohorts), but otherwise the R_0 matrix is unstructured. Each student's R_i contains only those rows and columns from R_0 that match the subjects and grades for which the student has test scores. In this way, the MRM is able to use all available scores from each student.

Algebraically, the district MRM is represented as:

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

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:

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

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

The MRM uses multiple years of student testing 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.

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 their prior and current testing data. The model produces means in each subject, grade, and year that can be used to calculate differences in order to obtain gains. Because students may change schools from one year to the next (in particular when transitioning from elementary to middle school, for example), the estimated mean score for the prior year/grade utilizes students that existed in the current year of that school. Therefore, mobility is taken into account within the model. Growth of students is computed using all students in each school including those that may have moved buildings from one year to the next.

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

Furthermore, in addition to reporting the estimated mean scores and mean gains produced by these models, the value-added reporting includes (1) cumulative gains across grades (for each subject and year), (2) multi-year up to 3-average gains (for each subject and grade), and (3) composite gains across subjects. These composites are explained in more detail in Section 6. In general, these are all different forms of linear combinations of the fixed effects, and their estimates and standard errors are computed in the same manner described above.

3.1.3.2 Teacher-level

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

- Students' performance with both their current and previous teacher effects are incorporated.
- Each teacher estimate accounts for multiple years of the students' testing data.
- The percentage of instructional responsibility the teacher has for each student is used.
- When next year's student scores are obtained, the previous year's teacher estimates can be refined with this additional information on student performance.

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

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

Algebraically, the teacher model is represented as:

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

y_{ijkl} is the test score for the i^{th} student in the j^{th} subject in the k^{th}grade in the l^{th} year. $\tau_{ijk^*l^*t}$ is the teacher effect of the t^{th} teacher on the i^{th} student in the j^{th} subject in grade k^* in year l^*. The complexity of the parenthesized term containing the teacher effects is due to two factors. First, in any given subject/grade/year, a student may have more than one teacher. The inner (rightmost) summation is over all the teachers of the i^{th} student in a particular subject/grade/year. $\tau_{ijk^*l^*t}$ is the effect of the t^{th} teacher. $w_{ijk^*l^*t}$ is the fraction of the i^{th} student's instructional time claimed by the t^{th} teacher. Second, as mentioned above, this model allows teacher effects to accumulate over time. That is, a teacher effect depends on students' performance in the current subject/grade/year as well as 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 state-level mean scores (for each year, subject and grade) in the fixed-effects vector b. 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 as a gain expressed as a deviation from the average gain for the state in a given year, subject, and grade.

Table 3 illustrates how the Z matrix is encoded for three students who have three different scenarios of teachers during grades three, four, and five in two subjects, math (M) and reading (R). Teachers are identified by the letters A–F.

Tommy's teachers represent the conventional scenario: Tommy is taught by a single teacher in both subjects each year (teachers A, C, and E in grades 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 C but also for third grade teacher A. This is how the "layering" is encoded. Similarly, in the grade 5 rows, there are ones for grade 5 teacher E, grade 4 teacher C, and grade 3 teacher A.

Susan is taught by two different teachers in grade 3: teacher A for math and teacher B for reading. In grade 4, Susan had teacher C for reading. 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 reading. This row is completely omitted from the Z matrix. There will always be a Z matrix row corresponding to each test score in the y vector. Since Susan has no entry in y for grade 5 reading, there can be no corresponding row in Z.

Eric's scenario illustrates team teaching. In grade 3 reading, Eric received an equal amount of instruction from teachers A and B. 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 E and F, they did not make an equal contribution. Teacher E claimed 80% responsibility, and teacher F claimed 20%.

Because teacher effects are treated as random effects in this approach, their estimates are obtained by shrinkage estimation, technically known as best linear unbiased prediction or as empirical Bayesian estimation. This means that *a priori* a teacher is considered "average" (with a teacher effect of zero) until there is sufficient student data to indicate otherwise. This method of estimation protects against false positives (teachers incorrectly evaluated as most effective or least effective), particularly in the case of teachers with few students so that random measurement error in the tests scores does not unduly affect their value-added measures.

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).

Similar to the district and school reporting, the teacher model provides estimated mean gains as well as (1) cumulative gains across grades (for each subject and year), (2) multi-year-average gains (for each subject and grade), and optionally (3) composite gains across subjects. All quantities can be described by linear combinations of the fixed and random effects and are found using the equations mentioned above.

Table 3: Encoding the Z matrix

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

3.1.4 Where the MRM is used in Ohio

The MRM is used with the OST in math and reading in grades 3–8. All data is used in each of the three separate analyses to obtain value-added measures at the district, school, and teacher level in grades 4–8.

In Ohio, multiple MRM analyses are run using the accountable district and school as well as the tested district and school information. For a detailed description of what is meant by accountable district and school in Ohio, see http://education.ohio.gov/getattachment/Topics/Data/Report-Card-Resources/Sections/Report-Card-Resources/WHERE-KIDS-COUNT.pdf.aspx.

The following analyses are done using the MRM methodology:

- Accountable district-level analyses
  - Overall
  - Gifted students
  - Lowest 20% of students
  - Students with disabilities
  - ESSA subgroups, including White, Black, Asian/Pacific Islander, American Indian, Multi-Racial, Hispanic, LEP, and ED
- Accountable school-level analyses
  - Overall
  - Gifted students
  - Lowest 20% of students
  - Students with disabilities
  - ESSA subgroups, including White, Black, Asian/Pacific Islander, American Indian, Multi-Racial, Hispanic, LEP, and ED
- Tested district-level analyses
  - Overall
- Tested school-level analyses
  - Overall
- STEM provider analysis
- Teacher-level analysis
- Dropout recovery analysis

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. For each subject, measures are also given across grades, across years (three-year averages), and combined across years and grades. In addition, composites of math and reading for each grade/year, across grades, across years (up to three-year averages), and across grades and years are computed for the different analyses. The composites for across years or across grades/years at the district and school level include both OST math and reading, even if one of those subjects does not have a value-added measure in the current (most recent analysis) year. Note that reporting based on the LEP subgroup analysis is not subject-specific.

At the teacher level, in addition to value-added measures for each OST subject/grade/year, a multi-year trend for each subject/grade for up to three years and a composite of math and reading across grades and years (up to two years) are also computed (and displayed on the EVAAS web application available at https://ohiova.sas.com/). The composite for teachers includes all prior value-added reporting, and this is a change from previous years where the teacher's composite included only the subjects for which the teacher had a value-added measure in the current (most recent analysis) year.

For more information about these composites and multi-year averages, see Section 6.

3.1.5 Students included in the analysis

All students are included into these analyses if they have scores that can be used. All available OST math and reading results for each student are incorporated into the models. Some student scores may be excluded if they are flagged as outliers or due to the other business rules described in Section 8.3. Because this model follows students from one grade to the next and measures growth through the movement from one grade to the next, the model assumes typical grade patterns for students. Students with non-traditional patterns, such as those who have been retained in a grade or skipped a grade, are treated as separate students in the model. In other words, these students are still included in the model, but the student is treated as separate students in different cohorts when these non-traditional patterns occur. This process occurs separately by subject since some students can be accelerated in one subject and not the other.

3.1.5.1 Overall accountable district- and school-level analysis

The analyses used to produce scores used for school and district report cards are all based on the business rules governing the accountability system. For more information on the "Full Academic Year/Where Kids Count Rules," see http://education.ohio.gov/getattachment/Topics/Data/Report-Card-Resources/Sections/Report-Card-Resources/WHERE-KIDS-COUNT.pdf.aspx.

For purposes of diagnostic interpretation, the EVAAS web application available to educators provides reports that are not based on the Accountability rules but only where students took their tests. In most cases, the "accountable" district and school are the same as the "tested" district and school. However, there are some cases where these are different. As an example, there could be students with disabilities who are held accountable to a different school or only the district level and not the school where they may have tested. There are also students who are accountable to the district or the state for various purposes.

3.1.5.2 Gifted district- and school-level analysis

The gifted student analysis pertains only to those students who are included in the "accountable student" set as described in 3.1.5.1. Students are included in the math analysis if they are either identified as gifted in math or superior cognitive. In the math analysis, students' prior and current math and reading test scores are included. Similarly, for reading, students are included who are identified as gifted in reading or superior cognitive. All other math and reading scores from those students are included in the reading analysis. Value-added measures are calculated for this subset of students for each district and school that meet the minimum requirements of student data. In this 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.

3.1.5.3 Students with disabilities district- and school-level analysis

The students with disabilities analysis pertains only to those students who are included in the "accountable student" set as described in 3.1.5.1. Students are included in the analysis if they are denoted as students with disabilities as recorded by the disability flag in EMIS. Value-added measures are calculated for this subset of students for each district and school that meet the minimum requirements of student data. In this 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.

3.1.5.4 Lowest 20% achievement district- and school-level analysis

The lowest 20% achievement student analysis pertains only to those students who are included in the "accountable student" set as described in 3.1.5.1. Students are included in the math analysis if the average of their current year/grade math score and prior year/grade math score is in the bottom 20% across the state. This bottom 20% is defined in the current (most recent analysis) year for each grade using the average of the current and prior year/grade scores. In the math analysis, these students' prior and current math and reading test scores are included. Similarly, for reading, students are included that are in the lowest 20% of statewide student achievement as defined above with the current and prior year/grade scores. All other math and reading scores from those students are included in the reading analysis. Value-added measures are calculated for this subset of students for each district and school that meet the minimum requirements of student data. In this 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.

For example, a student's grade 5 OST math score from last year and grade 6 OST math score from this year would be used to create his or her average math score. Similarly, the student's grade 5 OST reading score from last year and grade 6 OST reading score from this year would be used to create his or her average reading score. Students who do not have both scores in consecutive grades for a particular subject do not have an average and are not included. For each grade in a particular subject, the cut score is identified such that at least 20% of the students have an average score below that cut score. These are the students whose scores will be included in the value-added analysis for low achieving students for that subject. If a student's average math score is in the lowest 20% for math while his or her average reading score is not in the lowest 20% for reading, the value-added analysis for math will include both math and reading scores from the current and prior years. However, the student is not included in the analysis for reading. If a student is included in that subject, then the student's current year and prior year scores from math and reading are included in the modeling for that subject.

3.1.5.5 ESSA Accountability subgroup district- and school-level analysis

Ohio uses subgroup-level value-added measures in their federal accountability system. The subgroups include White, Black, Asian/Pacific Islander, American Indian, Multi-Racial, Hispanic, LEP, and ED. 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 OST subjects with a composite across math in grades 4–8 and reading in grades 4–8.

3.1.5.6 Community School Closure analyses

The community school closure analyses utilize all students that are accountable to that community school that have been at that same community school for at least two years in a row. If a student has been accountable to the school for the first time in a given year, then they are excluded from the analyses.

3.1.5.7 Teacher-level

The teacher value-added reports use all available math and reading test scores for each individual student linked to a teacher through the Ohio linkage roster verification process unless a student or a student test score meet certain criteria for exclusion.

Students are excluded from the teacher analysis if the students have more absences than an amount prescribed by law, which is currently set at 45 excused or unexcused days. (See **ORC 3319.112(A)(1)(b)**.) ODE provides EVAAS with a file that flags students who should be excluded based on that legislative action. Some student scores may also be excluded if they were flagged as outliers. (See Section 8.3.8.)

3.1.6 Minimum number of students for reporting

3.1.6.1 District and school level

To ensure 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 the subject/grade/year. This association could mean they were tested at the school or district or accountable to that school or district depending on what analysis is being conducted.
- 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 six students in each subject/year/grade in order for that subject/year/grade to be used in the gain calculation.

For example, to report an estimated NCE gain for school A in OST math grade 5 for this year, there must be the following requirements:

- There must be at least six fifth grade students with an OST math grade 5 score at school A for this year.
- At least one of the fifth-grade students at school A this year must have an OST math grade 5 score from this year *and* an OST math grade 4 score from last year.
- Of the fifth-grade students at school A this year *in all subjects, not just math*, there must be at least six students with an OST math grade 4 score from last year.

3.1.6.2 Teacher-level

The teacher-level 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 Year Equivalent (FYE) students in a specific subject/grade/year. The teacher's number of FYE 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 instance, if a teacher taught 10 students for 50% of their instructional time, then the teacher's FYE 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, then that teacher would have six FYE students and would receive a teacher value-added report. The instructional time attribution is obtained from the linkage roster verification process that is in use in Ohio. This information is in the files sent to EVAAS described in Section 2. 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 may 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 "gain," meaning the same subject prior test score must come from the immediate prior year and prior grade.

The process for creating an accurate link between students and teachers (roster verification) allows teachers and principals to review the attribution used in the EVAAS reports. For more information about teacher roster verification, see http://education.ohio.gov/Topics/Teaching/Educator-Evaluation-System/Ohio-s-Teacher-Evaluation-System/Student-Growth-Measures/Value-Added-Student-Growth-Measure/Value-Added-Roster-Verification.

3.1.7 Dropout recovery analysis

Growth measures are required for dropout recovery programs and, given the unique nature of student enrollment, student grade, and student testing in these programs, ODE has customized the value-added modeling and data inputs for a more meaningful growth measure. This analysis uses the school MRM methodology and is provided for the 85 schools and 15,000 students participating in these programs. It uses the same business rules described above, but there are a few additional business rules for this analysis that are described here.

3.1.7.1 Data inputs

At the dropout recovery programs, students participate in two testing windows during a given year: one at the beginning of the program and another at the end of the program/spring. For students who are enrolled at a dropout recovery program for more than one year, the model will use all available test scores while in the dropout recovery program.

The tests used in this analysis were selected by ODE for use in this project. One property of those assessments is that they are computer adaptive since the grade level can be difficult to determine for all students in the dropout recovery programs. More information about these assessments is here.

Students are only included in the model if they have a test in the beginning and a test at the end. These tests could include a math and/or reading assessment. The student only needs one subject at the beginning and one subject at the end to be included, and those could be different subjects. If the students have multiple scores in the same subject at the beginning or end of the year, then the lowest score would be used from the beginning, and the highest score would be used from the end of the year.

3.1.7.2 Modeling approach

The value-added model for dropout recovery programs is similar to the MRM described in this section. Modifications to the standard approach are described below.

As a first step, the distribution of scores for a subject/test window are mapped to a Normal Curve Equivalent distribution using the norm data provided by the test vendor. The average score for the first testing window of a specific program is compared to its average score for the second testing window. The growth standard (or expectation) is that students will maintain their achievement levels between the two testing windows relative to the norm referenced population, and the growth measure is the difference between the two achievement levels. As is the case with the OSTs, the MRM for dropout recovery programs provides the growth measure and its standard error for a particular test, and these will be used to calculate a growth index as described in Section 5 on page 32. As stated above, it is difficult to determine the grade of an individual student in these programs, so the normed reference group assumes tenth grade for all students in this analysis.

The difference in the interpretation is that the non-dropout recovery schools are measuring whether students maintained their same relative position in the distribution of statewide student achievement from one year to the next. The dropout recovery schools are using a national norm assessment and measuring whether students maintained their same relative position in that national norm referenced group from one year to the next.

3.2 Univariate Response Model (URM) for tests in non-consecutive grades

Tests that are not given for consecutive years require a different modeling approach from the MRM, and this modeling approach is called the univariate response model (URM). 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 all teachers were not administering a particular test in the state, then it would be compared to the average of those teachers with students taking that assessment.

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

- It does not require students to have all predictors or the same set of predictors, so long as a student has at least three prior test scores in any subject/grade.
- It minimizes the influence of measurement error by using multiple years of data for an individual student. Analyzing all subjects simultaneously increases the precision of the estimates.
- It allows educators to benefit from all tests, even when tests are on differing scales.

- It accommodates teaching scenarios where more than one teacher has responsibility for a student's learning in a specific subject/grade/year.

In Ohio, URM value-added reporting is available for the OST science tests in grades 5 and 8, OST social studies tests in grade 6 (previous years only), OST Algebra I, Mathematics I and II, Geometry, and ELA I and II at the district, school, and teacher levels. The end-of-course type tests that are used in the URM value-added reporting will include more subjects in the future when those subjects are administered to the majority of students who will be taking them as they are phased into the different graduating classes. Also, the URM methodology is also used in Ohio for other extended testing such as vendor assessments used in grades and subjects outside the state assessment scope.

3.2.1 URM at the conceptual level

The URM is run for each individual year, subject, and grade (if relevant). Consider all students who took grade 8 science in a given year. Those students are connected to all prior testing history (all grades, subjects, and years), and the relationship between the observed grade 8 science scores with all prior OST scores is examined. It is important to note that some prior test scores are going to have a greater relationship to the score in question than others. For instance, it is likely that prior science tests will have a greater relationship with science than prior reading scores. However, the other scores do still have a statistical relationship.

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

The measure of growth is a function of the difference between the observed (most recent) scaled scores and predicted scaled scores of students associated with each district, school, or teacher. If students at a school typically outperform their individual growth expectation, then that school will likely have a larger value-added measure. Zero is defined as the average district, school, or teacher in terms of the average progress, so 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 districts and schools 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.

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

In the case of team teaching, the single α_j is replaced by multiple αs, each multiplied by an appropriate weight, similar to the way this is handled in the teacher MRM in equation (13). The μ terms are means for the response and the predictor variables. α_j is the teacher effect for the j^{th} teacher—the teacher who claimed responsibility for the i^{th} student. The β terms are regression coefficients. Predictions to the response variable are made by using this equation with estimates for the unknown parameters (μs, βs, and 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:

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

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

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

Note that C in equation (16) is not the same as C in equation (4). This matrix is estimated using the Expectation Maximization algorithm for estimating covariance matrices in the presence of missing data provided by the Multiple Imputation procedure in SAS/STAT® (although no imputation is actually used). Only students who had a test score for the response variable in the most recent year and who had at least three predictor variables are included in the estimation. Given such a matrix, the vector of estimated regression coefficients for the projection equation (15) can be obtained as:

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

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

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

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

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

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

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

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

In the URM model, the effects (α_j) are considered random effects. Consequently, the $\hat{\alpha}_j$s are obtained by shrinkage estimation (empirical Bayes).[4] The regression coefficients for the ANCOVA model are given by the γs.

3.2.3 Students included in the analysis

The analyses that are used to produce scores used for school and district report cards are all based on the business rules governing the accountability system. For more information on the "Full Academic Year/Where Kids Count Rules," see http://education.ohio.gov/getattachment/Topics/Data/Report-Card-Resources/Sections/Report-Card-Resources/WHERE-KIDS-COUNT.pdf.aspx. These measures are only produced for OST Algebra I, Mathematics I, and ELA I at the district and school level for the report card system.

For purposes of diagnostic interpretation, the EVAAS web application available to educators provides reports that are not based on the Accountability rules but only where students took their tests. In most cases, the "accountable" district and school are the same as the "tested" district and school. However, there are some cases where these are different. As an example, there could be students with disabilities who are held accountable to a different school or only the district level and not the school where they may have tested. There are also students who are accountable to the district or the state for various purposes. These "tested" reports are produced for OST science tests in grades 5 and 8, OST social studies tests in grade 6 (previous years only), OST Algebra I, Mathematics I, and ELA I at the district, school, and teacher levels. OST Biology, American History, and American Government are only produced at the teacher level and not the district and school levels.

For a student's score to be used in the district-, school-, or teacher-level analysis for a particular subject/year and grade in cases of grade level tests, the student must have at least three valid predictor scores that can be used in the analysis, all of which cannot be deemed outliers. (See Section 8.3.8 on Outliers.) These scores can be from any year, subject, and grade that are used in the analysis. It will

[4] For more information on shrinkage estimation, 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, Second Edition* (Hoboken, NJ: John Wiley & Sons, 2008).

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. It is important to note that 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.

Unlike the MRM analysis, students with non-traditional grade patterns are included in the model as one student. Since this model does not determine growth based on consecutive grade movement on tests, students do not need to stay in one cohort from one year to the next. That said, if a student is retained and retakes the same test, then that prior score on the same test will not be used as a predictor in the URM for the same test as a response. This is mainly due to the fact that very few students used in the models have a prior score on the same test that could be used as a predictor.

3.2.4 Minimum number of students for reporting

3.2.4.1 District- and school-level

To receive a report, a tested 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.

3.2.4.2 Teacher-level

The teacher-level value-added *model* includes teachers who are linked to at least 10 students with a valid test score in the same subject and grade. To clarify, this means 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 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 Year Equivalent (FYE) students in a specific subject/grade/year. The teacher's number of FYE 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 instance, if a teacher taught 10 students for 50% of their instructional time, then the teacher's FYE number of students would be five, and the teacher would not receive a teacher value-added report. If another teacher taught twelve students for 50% of their instructional time, then that teacher would have six FYE students and that teacher would receive a teacher value-added report. The instructional time attribution is obtained from the linkage roster verification process that is used in Ohio. This information is in the files sent to EVAAS described in Section 2.

The process for creating an accurate link between students and teachers (roster verification) allows teachers and principals to review the attribution used in the EVAAS reports. For more information about teacher roster verification, see http://education.ohio.gov/Topics/Teaching/Educator-Evaluation-System/Ohio-s-Teacher-Evaluation-System/Student-Growth-Measures/Value-Added-Student-Growth-Measure/Value-Added-Roster-Verification.

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 OSTs.

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 definition of "expected growth" varies by model, and the precise definition depends on the selected model and state preference, and this section provides more details on the options and selections for defining expected growth. This document describes the expected growth as either a "base year" or "intra-year" approach. Base year refers to a growth expectation that is based on a particular year, say 2010, and any growth in the current year will be compared to the distribution of student scores in the base year. This is what was used in Ohio in previous years' reporting for math and reading in grades 4–8. Intra-year refers to a growth expectation that is always based on the current year (2015 for 2015 growth estimates, 2016 for 2016 growth estimates, and so on), and it has historically been used in Ohio for science and certain vendor assessments but is now used for all assessments including math and reading in grades 4–8.

4.1 Intra-year approach

4.1.1 Description

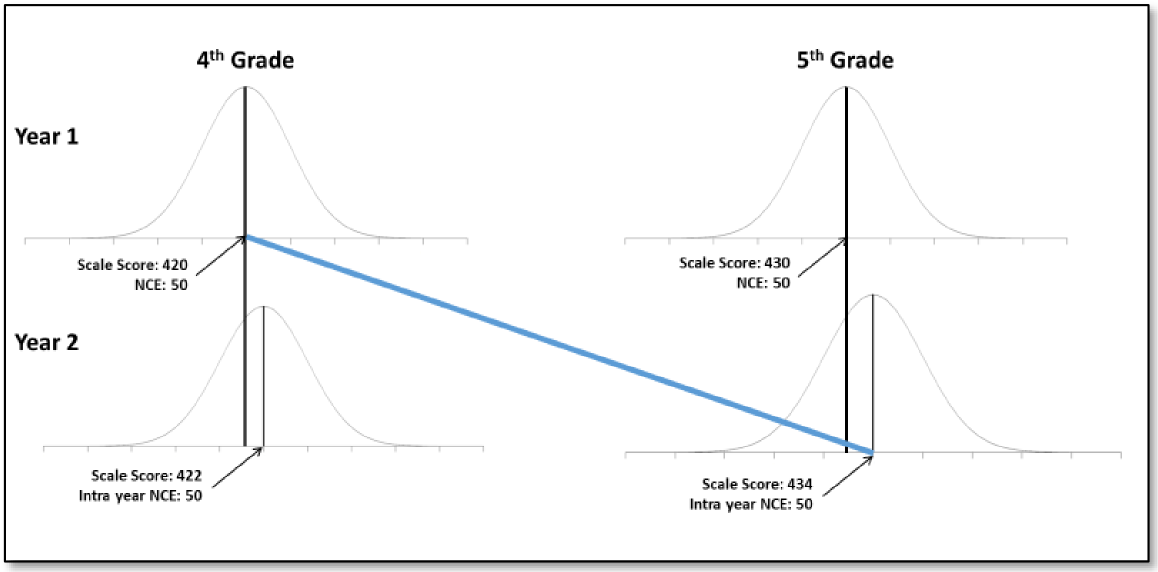
- Currently provided with URM reporting in science and certain vendor assessments in non-consecutive years and with MRM reporting in math and reading due to the transition to new assessments.
- URM definition: Students with a district, school, or teacher made the same amount of progress as students with the average district, school, or teacher in the state for that same year/subject/grade.
- MRM definition: Students maintained the same relative position with respect to the statewide student achievement that year.
- MRM simplified example: If students' achievement was at the 50th NCE in 2014 grade 4 math, based on the 2014 grade 4 math scale score distribution, and their achievement is at the 50th NCE in 2015 grade 5 math, based on the 2015 grade 5 math scale score distribution, then their estimated gain is 0.0 NCEs.
- Key feature: 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.

4.1.2 Illustrated example

Figure 3 below provides a simplified example of how growth is calculated with an intra-year approach when the state or pool achievement increases. The figure 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 grade 4 students in year 1 as they become grade 5 students in year 2. In year 1, our grade 4 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 5 distribution of scores in year 2*. The grade 5 distribution of scale scores in year 2 was higher than the grade 5 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 grade 4 in year 1 as they become grade 5 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.

Note 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 3: Intra-year approach example



4.2 Base year approach

4.2.1 Description

In years prior to 2014-15, the MRM value-added models used a “base year approach.” This means 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 simplified example, if students’ achievement was at the 50th NCE in 2010 grade 4 math, based on the 2010 grade 4 math scale score distribution, and at the 52nd NCE in 2011 grade 5, based on the 2010 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.

4.2.1.1 Stabilized NCE Scores

Even though standard psychometric methods are used to provide for equivalent scales within a grade and subject, it is recognized that unanticipated variability in the Ohio Achievement Assessment (OAA) scaling emerged across grades within a single year of testing, and across years within a grade. Therefore, in Ohio reporting prior to 2014-2015, the scale score distributions were converted into stabilized NCE scores using the statewide student achievement data in the base year (set then at 2010). The mapping from scale scores to NCEs was further modified with the "scale stabilization procedure" to compute the NCEs for each subject, grade, and year. The growth standard was given by maintaining the relative position in the statewide distribution of student achievement in the base year (2010) from grade to grade after stabilization. The scale stabilization procedure is described in detail at: http://education.ohio.gov/getattachment/Topics/Data/Accountability-Resources/Value-Added-Resources/OHIO-SCALE-STABILIZATION-FINAL-1.pdf.aspx.

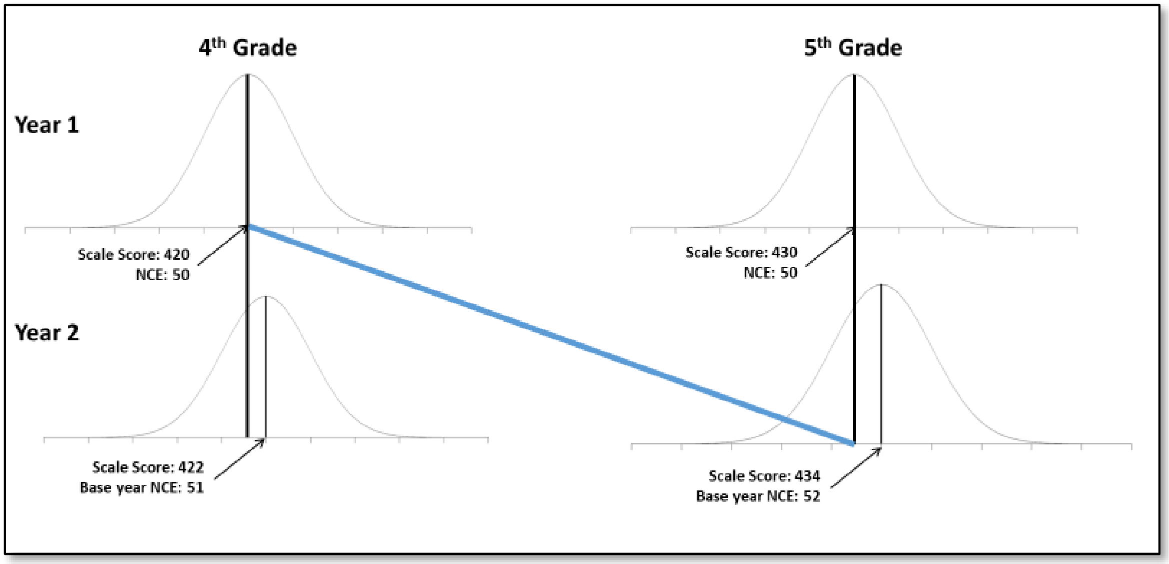
In general, when moving to a new assessment, as was the case in Ohio for the 2014-15 reporting, the intra-year approach can be used during the transition between old and new assessments. This will convert the scale scores of each of the different assessments to NCEs within each year. The growth standard expectation is then based on maintaining the same relative position with respect to all of a student's peers. This approach is useful when the assessment changes scales from one year to the next. The intra-year approach will be used for at least one additional year after the assessment change to ensure there has been a smooth transition. More details about the growth expectation of the intra-year approach are in Section 4.1.

4.2.2 Illustrated example

Figure 4 below provides a *simplified* example of how growth is calculated with a base year approach when the state achievement increases. The figure has four graphs, each of which plot the NCE distribution of scale scores for a given year and grade. In Ohio, the base year was most recently set at 2010, and the figure shows how the gain is calculated for a group of grade 4 students in year 1 as they became grade 5 students in year 2. In year 1, the grade 4 students scored, 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 scored, on average, 434 scale score points on the test, which corresponds to a 52nd NCE *based on the grade 5 distribution of scores in year 1*. The grade 5 distribution of scale scores in year 2 was higher than the grade 5 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 was required for students to make expected growth, which would be to maintain their position at the 50th NCE for grade 4 in year 1 as they became grade 5 students in year 2. The growth measure for these students was year 2 NCE – year 1 NCE, which would be 52 – 50 = 2. Similarly, if a group of students started out at the 35th NCE in grade 4 in year 1 and then moved their position to the 37th NCE in grade 5 in year 2, 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 4: 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.

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.

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

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

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 educators and policymakers with critical information regarding the certainty that students in a district, school or classroom are making decidedly more or less than the expected progress. Taking the standard error into account is particularly important for reducing the risk of misclassification (for example, identifying a teacher as ineffective when he or she is truly effective) for high-stakes usage of value-added reporting.

Furthermore, because the MRM and URM models utilize robust statistical approaches as well as maximize the use of students' testing history, they can provide value-added estimates for relatively small numbers of students. This allows more teachers, schools, and districts to receive their own value-added estimates, which is particularly useful to rural communities or small schools. As described in Section 3, there are minimum requirements between six and 10 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 to the value-added reporting.

5.2 Defining effectiveness in terms of standard errors

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

The standard error can help indicate whether a value-added estimate is significantly different from the growth standard. This growth standard is defined in different ways, but it is typically represented as zero on the growth scale and considered to be the *expected growth*. In the Ohio reporting, the value-added measures are placed in different categories based on the following:

- **Dark Green (Most Effective or "A")** is an indication that the growth measure is two standard errors or more above the growth standard (0). This level of certainty is significant evidence of exceeding the standard for academic growth.

- **Light Green (Above Average** or "B") is an indication that the growth measure is at least one but less than two standard errors above the growth standard (0). This is moderate evidence of exceeding the standard for academic growth.
- **Yellow (Average or "C")** is an indication that the growth measure is less than one standard error above the growth standard (0) and no more than one standard error below it (0). This is evidence of meeting the standard for academic growth.
- **Orange (Approaching Average or "D")** is an indication that the growth measure is more than one but no more than two standard errors below the growth standard (0). This is moderate evidence of not meeting the standard for academic growth.
- **Red (Least Effective or "F")** is an indication that the growth measure is more than two standard errors below the growth standard (0). This level of certainty is significant evidence of not meeting the standard for academic growth.

The terminology might be slightly different depending on what analysis is being categorized. For instance, teacher-level reporting uses the same boundary definitions, but the language is different to indicate the teacher-level analysis. 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 estimate or mean gain 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.***

5.3 Rounding and truncating rules

As described in the previous section, the effectiveness categories are based on the value of the growth index. As additional clarification, 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 provides the highest category given any type of rounding or truncating situation. For example, if the 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 Composite calculations

A composite combines value-added measures from different subjects, grades, and/or years. The sections below describe the calculation of composites for teacher reports, then schools, and lastly principals.

6.1 Teacher-level composites

The composite for teachers uses the most appropriate and robust statistical approach possible in the calculation of the value-added estimate and associated standard error. While the following text provides a specific example of a teacher's composite, the key policy decisions can be summarized as follows:

- The composite for teachers includes all prior value-added reporting (rather than only the subjects for which the teacher has a value-added measure in the *current* year as was the case in previous years). This change was made for URM with 2018 reporting and will be made with the MRM composite with the 2019 reporting.
- The composite for teachers weights each subject/grade based on the Full Year Equivalent number of students used in that measure within a year.

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

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

If a teacher does not have value-added measures from both the MRM and URM, then the composite index would be based on the model for which the teacher does have reporting. The following sections illustrate this process using value-added measures from a sample teacher, which are provided below:

Table 4: Sample Teacher Value-Added Information

Year	Subject	Grade	Value-Added Measure	Standard Error	Number of FYE Students
2017	Reading	8	-0.30	1.20	65
2017	Math	8	3.80	1.50	70
2018	Algebra I	8	11.75	6.20	20

6.1.1 Calculate MRM-based composite gain across subjects

All value-added measures from the MRM are in the same scale (Normal Curve Equivalents), so the composite gain across subjects is a simple average gain of all applicable gains with each weighted according to the proportion of students linked to that gain. For our sample teacher, the total number of FYE students affiliated with MRM value-added measures is 65 + 70, or 135. The reading grade 8 value-added measure would be weighted at 65/135 and the math grade 8 value-added measure would be weighted at 70/135.

More specifically, the sample teacher would have an MRM-based composite gain as follows:

$$MRM\ Comp\ Gain = \frac{65}{135}Read_8 + \frac{70}{135}Math_8 = \left(\frac{65}{135}\right)(-0.30) + \left(\frac{70}{135}\right)(3.80) = 1.83 \tag{20}$$

6.1.2 Calculate MRM-based standard error across subjects

6.1.2.1 Technical background on standard errors

As a reminder, the use of the word "error" does not indicate a mistake. Rather, value-added models produce *estimates*. That is, the value-added gains in the above tables are estimates, based on student test score data, of the teacher'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 or not 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 OST 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.

$$\begin{aligned} Var(aX + bY + cZ) &= a^2Var(X) + b^2Var(Y) + c^2Var(Z) \\ &+2ab\ Cov(X,Y) + 2ac\ Cov(X,Z) + 2bc\ Cov(Y,Z) \end{aligned} \tag{21}$$

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:

$$Cov(X,Y) = Correlation(X,Y)\sqrt{Var(X)}\sqrt{Var(Y)} \tag{22}$$

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.

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 teacher's composite gain, the relationship will generally be positive, and this means the MRM-based composite standard error is larger than it would be assuming independence.

6.1.2.2 Illustration of MRM-based standard error for sample teacher

For the sample teacher, it cannot be assumed that that the gains in the composite are independent because it is likely that some of the same students are represented in different value-added gains, such as grade 8 math in 2017 and grade 8 reading in 2017.

However, to demonstrate the impact of the covariance terms on the standard error, it is useful to calculate the standard error using (inappropriately) the assumption of independence. Using the MRM-based FYE weightings and standard errors reported in Table 4 and assuming total independence, the standard error would then be as follows:

$$\begin{aligned} MRM\ Comp\ SE &= \sqrt{\left(\frac{65}{135}\right)^2 (SE\ Read_8)^2 + \left(\frac{70}{135}\right)^2 (SE\ Math_8)^2} \\ &= \sqrt{\left(\frac{65}{135}\right)^2 (1.20)^2 + \left(\frac{70}{135}\right)^2 (1.50)^2} = 0.97 \end{aligned} \tag{23}$$

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 as follows using the MRM-based FYE weightings and standard errors from Table 4:

$$\begin{aligned} & MRM\ Comp\ SE \\ &= \sqrt{\left(\frac{65}{135}\right)^2 (SE\ Read_8)^2 + \left(\frac{70}{135}\right)^2 (SE\ Math_8)^2 + 2\left(\frac{65}{135}\right)\left(\frac{70}{135}\right)(SE\ Read_8)(SE\ Math_8)} \\ &= \sqrt{\left(\frac{65}{135}\right)^2 (1.20)^2 + \left(\frac{70}{135}\right)^2 (1.50)^2 + 2\left(\frac{65}{135}\right)\left(\frac{70}{135}\right)(1.20)(1.50)} = 1.36 \end{aligned} \tag{24}$$

The actual standard error will fall somewhere between the two extreme values of 0.97 and 1.36 with the specific value depending on the values of the correlations between pairs of value-added 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 example, if the 2017 grade 8 math and 2017 grade 8 reading classes had no students in common, then their correlation would be zero. On the other hand, if the 2017 grade 8 math and 2017 grade 8 reading classes contained many of the same students, there would be a positive correlation. However, even if those two classes had exactly the same students, the correlation would likely be considerably less than +1. Correlations of gains across years may be positive or slightly negative as the same student's score can be used in multiple gains. The actual correlations and covariances themselves are obtained as part of the EVAAS modeling process using equation (10) from Section 3.1.3. It would be impossible to obtain them outside of the modeling process. This process uses all the information about which students are in which subject/grade/year for each teacher. While this approach uses a more sophisticated technique, it more accurately captures the potential relationships among teacher estimates and student scores. This will lead to the appropriate standard error that will typically be between these two extremes, which are 0.97 and 1.36 in this example. In general, standard error of the composite gain will vary depending on the standard errors of the value-added gains and the correlations between pairs of value-added gains. The standard errors of the individual value-added gains will depend on the quantity and quality of the data that went into the gain, such as the number of students and the amount of missing data all those students have will contribute to the magnitude of the standard error.

6.1.3 Calculate MRM-based composite index across subjects

The final step is to calculate the MRM-based composite index, which is the composite value-added gain divided by its standard error. The composite index for the sample teacher is 1.83 divided by a number between 0.97 and 1.36. The actual MRM-based standard error is determined using all the information described above, which includes information beyond just our one sample teacher. For simplicity's sake, let's assume that the actual standard error in this example was 1.15, and the index for this teacher would be calculated as follows:

$$MRM\ Comp\ Index = \frac{MRM\ Comp\ Gain}{MRM\ Comp\ SE} = \frac{1.83}{1.15} = 1.59 \tag{25}$$

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

6.1.4 Calculate URM-based index across subjects

For our sample teacher (and for the majority of teachers who receive URM reporting in Ohio), 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. For the sample teacher, only a 2018 Algebra I growth measure is available.

$$URM\ Comp\ Index = \frac{Alg\ I\ VA\ Measure}{Alg\ I\ SE} = \frac{11.75}{6.20} = 1.90 \tag{26}$$

However, should a teacher 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 effective 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

6.1.5 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 linked to each model to determine the combined composite index. Our sample teacher has 155 students of which 135 are linked to the MRM and 20 to the URM, so 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:

$$Unadjusted\ Combined\ Comp\ Index = \left(\frac{135}{155}\right)(1.59) + \left(\frac{20}{155}\right)(1.90) = 1.62 \tag{27}$$

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 (1.62) 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:

$$Final\ Combined\ Comp\ SE = \sqrt{\left(\frac{135}{155}\right)^2 (1)^2 + \left(\frac{20}{155}\right)^2 (1)^2} = 0.88 \tag{28}$$

Therefore, the final combined composite index value is 1.62 divided by 0.88, or 1.85. This is the value that determines the teacher rating in the evaluation system.

6.2 District and school-level composites

Like the previous section, 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 policy decisions by ODE for schools can be summarized as follows:

- A composite is calculated for multiple subjects, grades, and years.
- A composite is calculated for a single year, up-to-two years, and up-to-three years of growth measures, and the web reporting will include the single year and up-to-three years composites.
- The composite for districts and schools may include OST math, reading, science, social studies, Algebra I, Mathematics I, Geometry, Mathematics II, ELA I, and ELA II.
- The composite for schools weights each subject/grade by the number of students in that subject/grade.

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 for a single-year composite using value-added measures from a sample middle school, which are provided below:

Table 5: 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	Reading	6	-1.10	1.00	46
2018	Math	7	2.00	0.50	50
2018	Reading	7	2.40	1.10	50
2018	Math	8	-0.30	0.60	40
2018	Reading	8	3.80	0.70	50
2018	Algebra I	N/A	-11.50	6.20	35

6.2.1 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 reading 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:

$$\begin{aligned} Comp\ Gain = & \frac{44}{280}Math_6 + \frac{46}{280}Read_6 + \frac{50}{280}Math_7 + \frac{50}{280}Read_7 + \frac{40}{280}Math_8 + \frac{50}{280}Read_8 \\ = & \left(\frac{44}{280}\right)(3.30) + \left(\frac{46}{280}\right)(-1.10) + \left(\frac{50}{280}\right)(2.00) + \left(\frac{50}{280}\right)(2.40) + \left(\frac{40}{280}\right)(-0.30) + \\ & \left(\frac{50}{280}\right)(3.80) = 1.76 \end{aligned} \tag{29}$$

6.2.2 Calculate MRM-based standard error across subjects

6.2.2.1 Technical background on standard errors

Similar to the teacher example, the standard error of the OST 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 reading in 2018. The statistical approach, outlined in Section 3.1.3 (with references), is quite sophisticated and will account for the correlations between pairs of value-added gains as shown in equation (21) and using equation (6) for schools and equation (10) for teachers.[5] 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.

6.2.2.2 Illustration of MRM-based standard error for sample school

As discussed in the teacher example, it cannot be assumed that the gains in the composite are independent because it is likely that some of the same students are represented in different value-added gains. Again, to demonstrate the impact of covariance terms on the standard error, it is useful to calculate the standard error using (inappropriately) the assumption of independence. Using the student weightings and standard errors reported in Table 5 and assuming total independence, the standard error would then be as follows:

[5] 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, Second Edition* (Hoboken, NJ: John Wiley & Sons, 2008).

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

$$= \sqrt{\begin{array}{l}\left(\frac{44}{280}\right)^2 (0.70)^2 + \left(\frac{46}{280}\right)^2 (1.00)^2 + \left(\frac{50}{280}\right)^2 (0.50)^2 \\ + \left(\frac{50}{280}\right)^2 (1.10)^2 + \left(\frac{40}{280}\right)^2 (0.60)^2 + \left(\frac{50}{280}\right)^2 (0.70)^2\end{array}} = 0.33 \qquad (30)$$

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, as was shown in the teacher level section.

The actual standard error will likely be above the value of 0.33 due to students being in both math and reading in the school with the specific value depending on the values of the correlations between pairs of value-added 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 that the actual standard error was 0.40 for the school composite in this example.

6.2.3 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:

$$MRM\ Comp\ Index = \frac{MRM\ Comp\ Gain}{MRM\ Comp\ SE} = \frac{1.76}{0.40} = 4.40 \qquad (31)$$

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

6.2.4 Calculate URM-based index across subjects

For our sample school (and for the majority of middle schools in Ohio), 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.

$$URM\ Comp\ Index = \frac{Alg\ I\ VA\ Measure}{Alg\ I\ SE} = \frac{-11.50}{6.20} = -1.85 \qquad (32)$$

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.

6.2.5 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, so 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:

$$Unadjusted\ Combined\ Comp\ Index = \left(\frac{280}{315}\right)4.40 + \left(\frac{35}{315}\right)(-1.85) = 3.71 \tag{33}$$

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:

$$Final\ Combined\ Comp\ SE = \sqrt{\left(\frac{280}{315}\right)^2 (1)^2 + \left(\frac{35}{315}\right)^2 (1)^2} = 0.90 \tag{34}$$

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 accountability overall grade. Different accountability measures use subsets of students, but the overall composite calculation is done the same.

6.2.6 Multi-year composites

The calculation for multi-year composites is the same as what was shown for the single-year composite. Any MRM growth measures could be combined across grades, subjects, and years within the model, and the model would provide the combined standard error as well. Any URM growth measures would be combined after they have been converted to growth indices as shown above using the index of the appropriate multi-year average as opposed to the single year value-added measure for a URM subject. Growth data from both models would be combined exactly as shown in the previous section.

6.3 Principal-level composites

This section captures how the policy decisions by ODE are implemented in the calculation of principal composites using school-level value-added data.

The key policy decisions and business rules for the principal-level composites can be summarized as follows:

- The term “principal” here refers to both assistant principals and principals. They are equivalent for the purposes of the calculations, and composites should be calculated for each person rather than per person per position.
- There are principals who fill the role of principal (P) for more than one school at a time.
- Many schools have more than one assistant principal (AP) at a time.

- Assume each school has a single principal at any given time until, after applying the business rules below, the data still show more than one principal at a school in a particular school year.
- Schools named “district testing” will be excluded.
- A principal (P) or assistant principal (AP) must be in the school for 120 school days (190 calendar days) of a single school year to be linked to a school for that year.
- The cutoff of a school year that is used for examining the data is from 9/15/year-1 to 5/31/year.
- If a principal or assistant principal starts in a school after 5/31/year and ends the position before 9/15/year, do not link the staff member to that school for that year.

6.3.1 Multiple principals reported at a school per year

The following steps describe the process to identify when there could be multiple principals reported a school in a given year:

1. Derive school years per principal per school from start and end dates of each principal based on information provided by ODE.
2. In cases where more than one principal is reported at a school with overlapping dates:
    a. Check the overlapping school years against the school year file provided by ODE to determine which school years personnel were reported as being employed as principals at the school. In that file, the SCHOOL_YEAR field shows the year in which a district reported the principal at that school.
    b. If a record of one of the overlapping principals per school does not show up in the school year file, exclude that record from the data used to compile the reports.
3. In cases where this approach does not narrow the data to a single principal per year (that is, the start/end dates overlap and there are more than one persons reported as principals in the same school per year), assume the school had more than one principal during that school year, and apply the value-added to both persons.
4. In cases where the school year file shows no principal reported for a school year, drop all overlapping records.

The same rules apply for assistant principals, and there are many more overlapping records.

6.3.2 Composite calculation

The following steps describe the policy decisions required to calculate the composite:

- Calculate a composite for each person. Treat assistant principal and principal positions as equivalent.
- A principal must be assigned to a school with a value-added measure in the most recent year to receive a composite. If the principal is not assigned to a school in the most recent year or the school to which he/she is assigned does not have a value-added composite, the principal will not receive a composite for that year.
- For the 2017-2018 reporting, the principal’s (and assistant principal’s) composites include growth measures based on a single year (the most recent year of reporting) and up-to-two years after applying above business rules.

- If a principal remains in the same school within a year, calculate the principal's single year estimate as the school composite across subjects and grades for that year.
- If a principal was in different schools within a year, calculate the principal's single year estimate (across schools) as the weighted average, adjusted for standard error, of the school composites for that year assuming independence. The weights are based on the number of subjects/grades in each school for that year.
- If the principal is in the same single school across two years, then the principal up-to-two year composite is the same up-to-two year composite for the school.
- If a principal is in more than one school across the two years, then individual year composites are calculated as described above for the two different scenarios of being in the same or different schools within a year. Those two individual year composites are then averaged weighted equally, and the standard error is adjusted assuming independence.

7 Projection model

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

OST projections are provided to a student's next two tested grade-level OST based on that student's most recent tested grade, such as projections to grades 6 and 7 for students who most recently tested in grade 5. EOC projections are provided for students as soon as they have at least three test scores in common with the students in the most recent tested cohort. ACT/SAT projections are provided to students who last tested in grades 6–11.

OST projections are made to the performance levels of Basic, Proficient, Accelerated, and Advanced, and the individual cut scores depend on each subject and grade. ACT/SAT projections will be provided to the following cut scores based on the performance of the 2017 (rather than 2018) cohort:

- SAT Evidence-Based Reading and Writing to OH Remediation-Free Benchmark of 480
- SAT Mathematics to OH Remediation-Free Benchmark of 530
- ACT English to OH Remediation-Free Benchmark 18
- ACT Mathematics to OH Remediation-Free Benchmark of 22
- ACT Reading to OH Remediation-Free of Benchmark 22

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 (xs) are scores on tests the student has already taken, and the categorical variable is the school at which the student received instruction in the subject/grade/year of the response variable (y). Algebraically, the model can be represented as follows for the i^{th} student.

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i \tag{35}$$

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, βs, sometimes α_j). The parameter estimates (denoted with "hats," e.g., $\hat{\mu}$, $\hat{\beta}$) are obtained using the most current data for which response values are available. The resulting projection equation for the i^{th} student is

$$\hat{y}_i = \hat{\mu}_y \pm \hat{\alpha}_j + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots + \epsilon_i \tag{36}$$

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

Two difficulties must be addressed 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. In addition to the projected score itself, the standard error of the projection is calculated ($SE\,(\hat{y}_i)$). Given a projected score and its standard error, it is possible to calculate the probability that a student will reach some specified benchmark of interest (b). Examples are the probability of scoring at the proficient (or advanced) level on a future end-of-grade test or the probability of scoring sufficiently well on a college entrance exam to gain admittance into a desired program. The probability is calculated as the area above the benchmark cutoff score using a normal distribution with its mean equal to the projected score and its standard deviation equal to the standard error of the projected score as described below. Φ represents the standard normal cumulative distribution function.

$$Prob(\hat{y}_i \geq b) = \ \Phi\left(\frac{\hat{y}_i - b}{SE(\hat{y}_i)}\right) \qquad (37)$$

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 teacher credential ID only as requested by ODE because other information, such as teacher name, may change over time, but credential ID remains the same.

8.2 Checks of scaled score distributions

The statewide distribution of scale scores is examined each year to determine if 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 EVAAS. 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 students’ growth for students who would have otherwise scored higher or lower than the test allowed. It is also important there are enough test scores at the high or low end of achievement, so measurable differences can be observed. Stretch can be determined by the percentage of students who score near the minimum or the maximum level for each assessment. In 2015, the percentage of students who achieved a maximum score on the OST end-of-grade and end-of-course assessments ranged from a high of 0.38% (sixth-grade social studies) to a low of .01% (seventh-grade math). As an example, if a much larger percentage of students scored at the maximum in one grade than in the prior grade, then it may seem that these students had negative growth at the very top of the scale when it is likely due to the artificial ceiling of the assessment. Percentages for all OST assessments are well below acceptable values, meaning that the OSTs 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 is aligned with the curriculum. The requirement that tested material correlates with standards will be met if the assessments are designed to assess what students are expected to know and be able to do at each grade level. Since the OSTs are designed to measure state curriculum, this is not an issue.

8.2.3 Reliability

Reliability can be viewed in a few different ways for assessments. Psychometrics view reliability as the idea that a student would receive similar scores if the assessment was taken multiple times. Reliability also refers to the assessment's scales across years; both types of reliability are important when measuring growth. The first type reliability is important for most any use of standardized assessments. The second type of reliability is important when a base year is used to set the expectation of growth since this approach assumes that scale scores mean the same thing in a given subject and grade across years. Starting with the 2014-15 reporting, the intra-year approach will be used, so this is less of a concern.

8.3 Data quality business rules

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

8.3.1 Missing grade levels

In Ohio, 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 End-of-Grade type tests, then these records will be excluded from all analyses. The grade is required to include a student's score in the appropriate part of the models, and it would need to be known if the score was to be converted into an NCE.

Of the 1,863,071 records from the 2017-18 OST Math, Reading, and Science assessments, no records were excluded due to this business rule.

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 accountable school, then extra scores will be excluded from the analysis and reporting.

Of the 2,771,146 records from the 2017-18 OST Math, Reading, Science, Algebra I, Geometry, ELA I, ELA II, Mathematics I, Mathematics II, Biology, American History, and American Government assessments, no records were excluded due to this business rule.

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

If a student has a score with a missing accountable district or school for a particular subject and grade in a given testing period, then the duplicate score that has an accountable district and/or school will be included over the score that has the missing data.

Of the 2,771,146 records from the 2017-18 OST Math, Reading, Science, Algebra I, Geometry, ELA I, ELA II, Mathematics I, Mathematics II, Biology, American History and American Government assessments, 642 records (0.02%) were excluded due to this business rule.

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 accountable schools, then both scores will be excluded from the analysis.

Of the 2,771,146 records from the 2017-18 OST Math, Reading, Science, Algebra I, Geometry, ELA I, ELA II, Mathematics I, Mathematics II, Biology, American History, and American Government assessments, 3,044 records (0.11%) were excluded due to this business rule.

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 if 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.

Of the 2,771,146 records from the 2016-17 OST Math, Reading, Science, Algebra I, Geometry, ELA I, ELA II, Mathematics I, Mathematics II, Biology, American History, and American Government assessments, 8 records (0.0003%) were excluded due to this business rule.

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 in 2017 to ninth in 2018) or is moved back by one grade or more (i.e. moves from fourth in 2017 to third in 2018) 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. In Ohio for the ODE analysis, EVAAS does not remove students with scores that appear to be associated with inconsistent grades. EVAAS leaves students in the analysis at the tested grade that EVAAS receives from ODE.

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

If a student is tested at two different accountable 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. In Ohio, it can happen that a student is accelerated in a subject and does test at two different accountable schools.

8.3.8 Outliers

Student assessment scores are checked each year to determine if they are outliers in context with all the 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 OST math grades are examined simultaneously, and any scores that appear inconsistent, given the other scores for the student, are flagged. 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 EVAAS web application.

This process is part of a data quality procedure to ensure 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 if 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. 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 -3.5 when looking at the difference between the score in question and the reference group of scores.
- 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 4.0.
- The percentile of the comparison score must be below a certain value.
- There must be at least 3 scores in the comparison score average.

Of the 2,771,146 records from the 2017-18 OST Math, Reading, Science, Algebra I, Geometry, ELA I, ELA II, Mathematics I, Mathematics II, Biology, American History, and American Government assessments, 691 records (0.03%) were excluded due to this business rule.

8.4 Teacher-student linkages

Student linkages are not used in the analysis if they are listed as having more than 45 unexcused absences. These linkages are excluded first. Of the 3,097,359 linkages from the 2017-18 OST Math, Reading, Science, Algebra I, Geometry, ELA I, ELA II, Mathematics I, Mathematics II, Biology, American History, and American Government assessments, 51,135 linkages (1.65%) were excluded due to this business rule.

Teacher student linkages are connected to assessment data based on the subject and identification information described above. There are some instances where extra processing is required for analysis. The value-added models place a restriction on how teachers can claim students, such that a student cannot be claimed by teachers more than 100%. Therefore, if a student is claimed in an individual year, subject, and grade at more than 100%, then the individual teacher's weight is divided by the total sum of all weights to redistribute the attribution of the student's test scores across teachers. A student can be claimed less than 100% for various reasons, so under-claimed linkages for a student are not modified.