Ohio Department of Education

Technical Documentation of EVAAS Analyses

Version 1.3 8 December 2014

Contents

1 Introduction

1.1 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 all prior achievement/prior results, with achievement being measured by a quality assessment such as the OAA tests.

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

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

The purpose of this document is to guide you through the value-added modeling *based on the statistical approaches, policies, and practices selected by the State of Ohio and currently implemented by SAS*. This document describes the input data, modeling and business rules for the district, school and teacher value-added reporting in Ohio.

The State of Ohio and the SAS team have provided value-added reporting since 2002. The initial collaboration was through Project SOAR, a 42-district pilot. By 2006, the 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 OAA assessments in grades four through eight 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

In order 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 that progress can be measured for both low-achieving students as well as high-achieving students. A floor or ceiling in the scales could disadvantage educators serving either low-achieving or high-achieving students.
- **The test is 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 SAS receives from Ohio.

The current value-added implementation includes assessments measuring Ohio's standards (OAA) and 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 is currently moving towards implementing new assessments. Changes in testing regimes occur at regular intervals within any state, and these changes need not disrupt the continuity and use of value-added reporting by educators and policymakers. Based on twenty years of experience with providing value-added and growth reporting to educators, SAS has developed several ways to accommodate changes in testing regimes.

Prior to any value-added analyses with new tests, SAS verifies that the test's scaling properties are suitable for such reporting. In addition to the criteria listed above, SAS 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 six math exam should be correlated to previous math scores in grades four and five 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

The OAA tests are administered in the spring semester except for OAA reading in grade three, which is given in the fall and/or spring semesters. In grade three, 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.

2.2.1 Tests given in consecutive grades for the same subject

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

- Ohio Achievement Assessment (OAA) mathematics in grades three through eight.
- OAA reading in grades three through eight.

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

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

- OAA science in grades five and eight.
- Ohio Graduate Test (OGT) in grade ten for mathematics, reading, science, social studies, and writing.

2.2.3 Student identification information

Ohio's state law prohibits ODE from maintaining student names; therefore, the data ODE sends to SAS contains only the state secure ID (SSID) for each student and no name information. IBM securely transfers student names to Battelle for Kids (BFK) and The Management Council of the Ohio Education Computer Network (MCOECN)**,** and those student names are matched using SSID and forwarded to SAS. 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 secure ID (SSID)

2.2.4 Assessment information provided

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

- 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

2.3 Student-level information

Student-level information is used in creating the web application to assist educators analyze the data to inform practice and assist all students with academic progress. This information is also used to create accountability categories that are reported to the public. SAS 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 as follows:

- Gifted-reading
- Gifted-math
- Superior cognitive
- Migrant
- Limited English Proficient
- Economically disadvantaged
- Students with disabilities
- Gender
- Race
    - American Indian/Alaskan Native
    - Asian/Pacific Islander (prior to 2011)
    - Asian (beginning in 2011)
    - Native Hawaiian or Other Pacific Islander(beginning in 2011)
    - Black, Non-Hispanic
    - Hispanic
    - White, Non-Hispanic
    - Multi-Racial

More information can be found in Ohio's 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 makes teacher-level analyses much more reliable and accurate.

For the purposes of Ohio's teacher-level value-added reports, SAS receives teacher identification data and student-teacher linkages from both BFK and MCOECN. The roster verification process provides the 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 the roster verification process, the teacher student linkage reported and verified through EMIS is sent and used by SAS.

2.5 Principal-level information

SAS receives data on individual principals and assistant principals linking each of them to their schools. SAS and ODE are currently discussing how school value-added data can be used along with this information to provide value-added data for principals and assistant principals.

2.6 Data files by source

Table 1: Data Files Received by SAS

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 all prior achievement/prior results, with achievement being measured by a quality assessment such as the OAA tests.

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, SAS 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 the OAA math and reading assessments in grades three through eight.
- **Univariate Response Model (URM)** is used when a test is given in non-consecutive grades, such as OAA science assessments in grades five and eight or any End-of-Course tests that may exist in the future.

Both models offer the following advantages:

- The models include all of each student's testing history without imputing any test scores.
- The models can accommodate students with missing test scores.
- The models can accommodate team teaching or other shared instructional practices.
- The models use up to five 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 all available test scores and including students, even if they have missing test scores, it is not necessary to make *direct* adjustments for students' background characteristics. In short, 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, K (Winter 2004). The Real Value of Teachers: If Teachers Matter, Why Don't We Act Like It? (The Education Trust: Washington DC).

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 (MRM) reporting for tests in consecutive grades

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

The MRM is a *gain-based model*, which means that it measures growth between two points in time for a group of students. The 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 the base year (currently 2010) for a specific subject and grade.

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

- All students with valid data are included in the analyses, even if they have missing test scores. All of each student's testing history is included without imputing any test scores.
- By including all students in the analyses, even those with a sporadic testing history, it provides the most realistic estimate of achievement available.
- It minimizes the influence of measurement error inherent in academic assessments by using multiple data points of student test history (up to five years of data for an individual student).
- 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.
- Using a state base year to benchmark growth, the MRM provides a means by which all LEAs and teachers can meet or exceed the growth standard (expectation). Additionally, a state base year accounts for changes in statewide progress over time. For more information on the base year and growth expectation, see Section 4.
- The model analyzes all subjects simultaneously to improve precision and reliability.

As a result 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, D. F., Han, B. and Lockwood, J. R. (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 D.C.
- On the **advantages of the longitudinal, mixed model approach**: Lockwood, J.R. and McCaffrey, D.F. (2007). "Controlling for Individual Heterogeneity in Longitudinal Models, with Applications to Student Achievement." Electronic Journal of Statistics, Vol. 1, 223-252.
- On the **insufficiency of simple value-added models**: McCaffrey, D. F., Han, B. and Lockwood, J. R. (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 in the base year for a specific subject and grade?

3.1.1 MRM at the conceptual level

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

Table 2: Scores without missing data

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

Table 3: Scores with missing data

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

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

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

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 conceptual idea of what is done with the school and district models. The teacher model is slightly more complex, and all models are explained in more detail below (in Section 3.1.3). The first step in the MRM is to define the scores that will be used in the model.

3.1.2 Normal curve equivalents

3.1.2.1 *Why SAS 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. For example, 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 0 or even negative.

The NCEs used in SAS 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 (within-year approach). The base year currently used in the Ohio MRM analysis is 2010, although the within-year approach is required when there is a change in testing regime and 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 *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 OAA scaling emerges across grades within a single year of testing, and across years within a grade. Therefore, in Ohio, the scale score distributions are converted into stabilized NCE scores using the statewide student achievement data in the base year (currently 2010). The

mapping from scale scores to NCEs is further modified with the "scale stabilization procedure" to compute the NCEs for each subject, grade, and year. The growth standard is given by maintaining the relative position in the distribution of the base year (2010) statewide distribution of student achievement 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 will be the case in Ohio, the within-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 very useful when the assessment changes scales from one year to the next. The within year approach will be used for at least one additional year after the assessment change to ensure that there has been a smooth transition. The within year approach is not in use at this time. More details about the growth expectation of the within-year approach are in Section 4.2.

3.1.2.3 *How SAS uses normal curve equivalents in MRM*

There are multiple ways of creating NCEs. SAS 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 4 provides an example of the way that SAS converts scale scores to NCEs.

The first five columns of Table 4 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 4 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 4. 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 4. 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 4.

Table 4: Converting tabulated test scores to NCE values

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

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

3.1.3 Technical description of the linear mixed model and the MRM

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

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

y (in the 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 (usually up to five 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:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.1.3.1 *District and school level*

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

$$y_{ijkl} = \mu_{ijkld} + \epsilon_{ijkld} \quad (11)$$

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:

$$y_{ijkls} = \mu_{ijkls} + \epsilon_{ijkls} \quad (12)$$

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

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

Solving the mixed model equations for the district or school MRM produces a vector b that contains the estimated mean score for each school (in the school model), subject, grade and year. To obtain a value-added measure of average student growth, a series of computations can be done using the students from a school in a particular year and all of their prior, current, and future 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 so that growth of students is computed using all students in each school including those that may have moved buildings from one year to the next.

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

Furthermore, in addition to reporting the estimated mean scores and mean gains produced by these models, the value-added reporting 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:

- The current and previous teacher effects are incorporated.
- Each teacher estimate takes into account all the students' testing data over the years,
- The percentage of instructional responsibility the teacher has for each student is used,
- The impact of previous teachers on the current year students and adjustments for future teachers is included (meaning, when next year's student scores are obtained, the previous year's teacher estimates can be refined with this additional information).

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

To allow for the possibility of many teachers with relatively few students per teacher, MRM enters teachers as random effects via the Z matrix in the linear mixed model. The X matrix contains a column for each subject/grade/year, and the b vector contains an estimated state 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:

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

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

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

Table 5 illustrates how the Z matrix is encoded for three students who have three different scenarios of teachers during grades three, four, and five in two subjects, math (M) and reading (R). Teachers are identified by the letters A–F.

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

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

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

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

From the computational perspective, the teacher gain can be defined as a linear combination of both fixed effects and random effects and is estimated by the model using equation (9). The variance and standard error can be found using equation (10).

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 of these quantities can be described by linear combinations of the fixed and random effects and are found using the equations mentioned above.

Table 5: 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 OAA test in math and reading in grades three through eight. All of this data is used in each of the three separate analyses to obtain value-added measures at the district, school, and teacher level in grades four through eight.

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/2012-2013-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

- Accountable school-level analyses
    - Overall
    - Gifted students
    - Lowest 20% of students
    - Students with disabilities
- Tested district-level analyses
    - Overall
- Tested school-level analyses
    - Overall
- Teacher-level 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), as well as 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), as well as 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 includes both OAA math and reading, even if one of those subjects does not have a value-added measure in the current (most recent analysis) year.

At the teacher level, in addition to value-added measures for each OAA 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 three years) are also computed (and displayed on the EVAAS web application available at https://ohiova.sas.com/). The composite for teachers includes only the subjects for which the teacher has 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 of every student's math and reading OAA results for the most recent five years 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.

3.1.5.1 *Overall accountable district and school level*

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/2012-2013-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, all of these students' prior tests and other tests in the same year are included. Similarly for reading, students are included who are identified as gifted in reading or superior cognitive. All other 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.

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.

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, all of these students' prior tests and other tests in the same year 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 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.

For example, a student's 2011 5th grade OAA math and 2012 6th grade OAA math scores would be used to create his or her average math score. Similarly, the student's 2011 5th grade OAA reading and 2012 6th grade OAA reading scores 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 all of the student's current year and prior year scores (even from other subjects) are included in the modeling for that subject.

3.1.5.5 *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.

3.1.5.6 *Teacher-level*

The teacher value-added reports use all available 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 SAS 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 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 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 2013 OAA math grade five, there must be the following requirements:

- There must be at least six fifth grade students with a 2013 OAA math grade five score at school A.
- At least one of the 2013 fifth grade students at school A must have a 2013 OAA math grade five score *and* a 2012 OAA math grade four score.
- Of the 2013 fifth grade students at school A *in all subjects, not just math*, there must be at least six students with a 2012 OAA math grade four score.

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, in order 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 ten 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 in use in Ohio. This information is in the files sent to SAS 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).

Students are linked to a teacher based on the subject area taught and the assessment taken. In some cases, the course being taught may not align to the assessment being taken and in those cases linkage is not mandatory. For example, all grade eight students take the OAA mathematics test. However, some of these eighth grade students are enrolled in Algebra I while the rest are enrolled in general grade eight math classes. If a teacher teaches Algebra I, then that teacher will not be automatically linked to those students and may not receive a math grade eight report. Districts MAY choose to have their teachers link to such students if they would like for them to be included in the teacher's report. If a teacher teaches both eighth grade general mathematics and Algebra I and the teacher is participating in extended testing, then that teacher would receive a grade eight mathematics report based on the students that took general math and an Algebra I report through extended testing.

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.

3.2 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 not all teachers were administering a particular test in the state, then it would 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 up to five 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 OAA science tests in grades five and eight at the district and school levels. It is not used at the individual teacher level because OAA science tests assess up to three years of academic content standards material rather than what is covered in the tested grade. In other words,

OAA science in grade five covers material from grades three, four and five, so the value-added reporting should not only be attributed to the grade five science teacher. 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 eight science in a given year. Those students are connected to all of their prior testing history (all grades, subjects, and years), and the relationship between the observed grade eight science scores with all prior OAA test scores is examined. It is important to note that some prior test scores are going to have a greater relationship to the score in question than others. For instance, it is likely that prior science tests will have a greater relationship with science than prior reading scores. However, the other scores do still have a statistical relationship.

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

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

3.2.2 Technical description of the district, school and teacher models

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

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

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

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

observed value for the specific response and have three predictor scores. The resulting prediction equation for the i^{th} student is as follows:

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

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,

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

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

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

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

The prediction equation also requires estimated mean scores for the response and for each predictor (the $\hat{\mu}$ terms in the prediction equation). These are not simply the grand mean scores. It can be shown that in an ANCOVA, if 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 takes into account missing data. The overall means ($\hat{\mu}$ terms) are then obtained as the simple average of the teacher-level means.

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

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

The $\hat{y}_i$ term is nothing more than a composite of all the student's past scores. It is a one-number summary of the student's level of achievement prior to the current year. The different prior test scores making up this composite are given different weights (by the regression coefficients, the $\hat{\beta}$'s) 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.

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

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

3.2.3 Students included in the analysis

The district and school that receive reporting using this analysis are the "tested" district and school associated with a student. There is no reporting done with the "accountable" district or school since this model is not used in the subjects and grades that are used in state accountability.

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

3.2.4 Minimum number of students for reporting

To receive a report, a tested district or school must have at least ten 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.

At this time, this model is not being used for any of the teacher level analyses for OAA, although many districts within the state receive teacher-level reporting based on the URM approach for other vendor assessments.

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 OAA tests

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

However, the 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 "within-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 currently what is used in the Ohio for math and reading in grades four through eight. Within-year refers to a growth expectation that is always based on the current year (2012 for 2012 growth estimates, 2013 for 2013 growth estimates, and so on).

Currently, Ohio uses a base year approach with math and reading and a within-year approach for science and certain vendor assessments. As described in greater detail below, there will be a need to switch to the within-year approach for all tested subjects and grades when the assessments change to PARCC.

4.1 Base year approach

4.1.1 Description

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

As a simplified example, if students' achievement was at the 50^{th} NCE in 2010 grade four math, based on the 2010 grade four math scale score distribution, and at the 52^{nd} NCE in 2011 grade five, based on the 2010 grade five math scale score distribution, then their estimated mean gain is 2 NCEs.

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

Following the implementation of any new assessments and changes in academic standards, it is best that the base year be reset to a within-year approach in order to accommodate the differences between the old and new testing regimes and minimize any impact on the value-added reporting. To be more specific, it is required to use the within-year approach if there is no mapping from the old assessment's scale to the new assessment's scale. However, even if that mapping does exist, the within-year approach is preferred so that there are not any unusual swings in value-added measures. If a base year approach is desired after the transition, SAS recommends that the new base year be selected after, at a minimum, the second year of the new assessment to verify the smooth transition. As a result, during the transition to new assessments in Ohio, SAS will use a within-year approach.

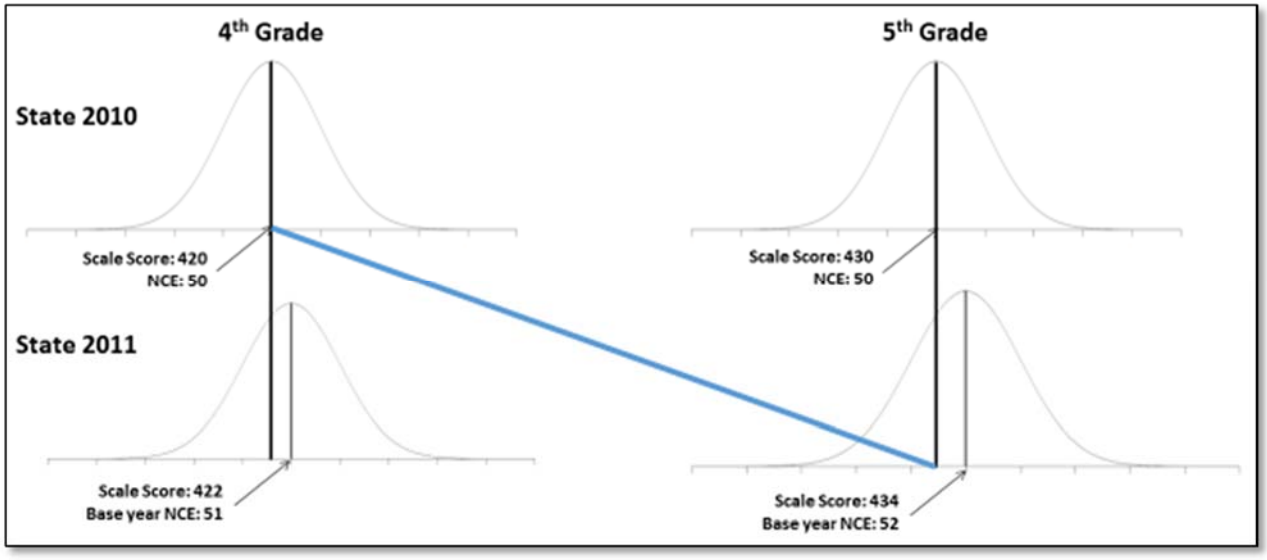
4.1.2 Illustrated example

The graphic below (Graph 1) provides a *simplified* example of how growth is calculated with a base year approach when the state achievement increases. The graphic below has four graphs, each of which plot the

NCE distribution of scale scores for a given year and grade. In Ohio, the base year is currently 2010, and the graphic shows how the gain is calculated for a group of 2010 grade four students as they become 2011 grade five students. In 2010, our grade four students score, on average, 420 scale score points on the test, which corresponds to the 50th NCE (similar to the 50th percentile). In 2011, the students score, on average, 434 scale score points on the test, which corresponds to a 52nd NCE *based on the 2010 grade five distribution of scores*. The 2011 grade five distribution of scale scores was higher than the 2010 grade five distribution of scale scores, which is why the lower right-hand graph is shifted slightly to the right. The blue line shows what is required for students to make expected growth, which would to maintain their position at the 50th NCE in 2010 grade four as they become 2011 grade five students. The growth measure for these students is 2011 NCE – 2010 NCE, which would be 52 – 50 = 2. Similarly, if a group of students started out at the 35th NCE in 2010 grade four and then moved their position to the 37th NCE in 2011 grade five, they would have a gain of two NCEs as well.

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

Graph 1: Base year approach example



4.2 Within-year approach

4.2.1 Description

- Currently provided with URM reporting in science and certain vendor assessments in non-consecutive years
- Will be used in the MRM reporting during 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 2010 grade four math, based on the 2010 grade four math scale score distribution, and their achievement is at the 50th NCE in 2011 grade five math, based on the 2011 grade five math scale score distribution, then their estimated gain is 0.0 NCEs.

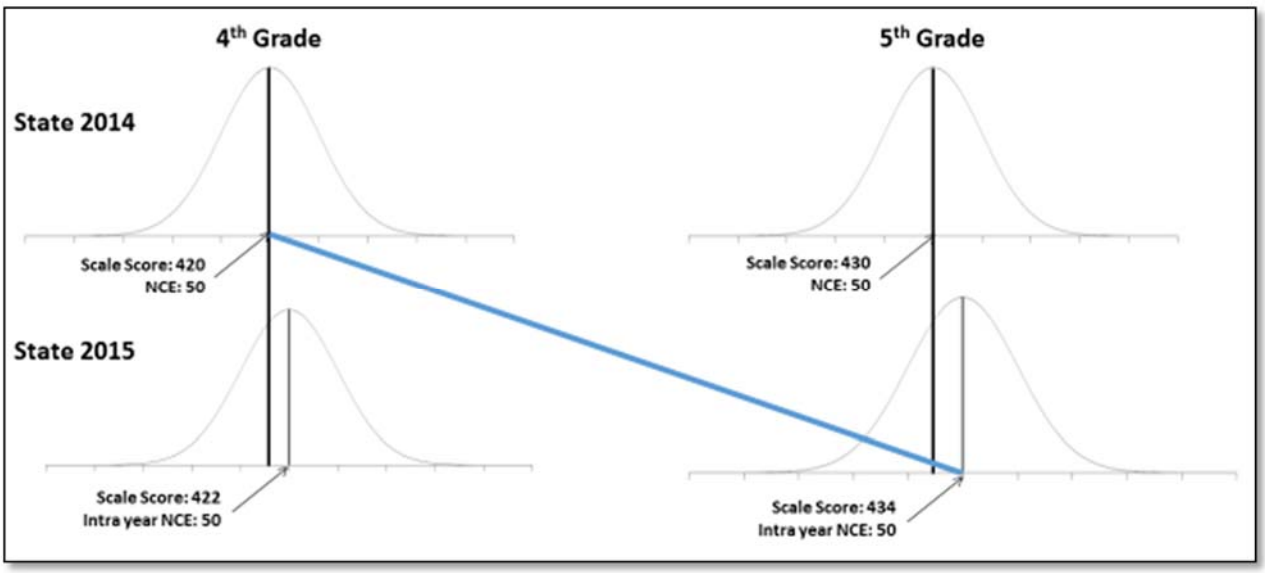
- Key feature: The value-added measures tend to be centered on the growth expectation every year, with approximately half of the district/school/teacher estimates above zero and approximately half of the district/school/teacher estimates below zero.

4.2.2 Illustrated example

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

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

Graph 2: Within-year approach example



4.3 Defining the expectation of growth during an assessment change

During the change of assessments, the scales from one year to the next will be completely different from one another. This does not present any particular changes with the URM methodology because all predictors in this approach are already on different scales from the response variable, so the transition is no different from a

scaling perspective. Of course, there will be a need for the predictors to be adequately related to the response variable of the new assessment, but that typically is not an issue.

However, with the MRM methodology, a base year approach presents challenges since it requires the scales to stay consistent over time. That said, with the within-year approach, 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, SAS 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 the educators and policymakers with critical information regarding the certainty that students in a district, school or classroom are making decidedly more or less than the expected progress. Taking the standard error into account is particularly important for reducing the risk of misclassification (for example, identifying a teacher as ineffective when he or she is truly effective) for high-stakes usage of value-added reporting.

Furthermore, because the MRM and URM models utilize robust statistical approaches as well as maximize the use of students' testing history, they can provide value-added estimates for relatively small numbers of students. This allows more teachers, schools, and districts to receive their own value-added estimates, which is particularly useful to rural communities or small schools. 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, the teachers may have students with very different amounts of prior testing history. Due to this variation, the standard errors in a given subject/grade/year could vary significantly among teachers, depending on the available data that is associated with their students, and it is another important protection for districts, schools and teachers to incorporate standard errors 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 that are 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.***

The distribution of these categories can vary by year/subject/grade. There are many reasons this is possible, but overall, it can be shown that there are more measurable differences in some subjects and grades compared to others.

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 Multi-year and composite calculations

The first part of this section captures how the policy decisions by ODE are implemented in the calculation of composites and multi-year trends for teachers in the tested subjects and/or grades. The last part of this section summarizes the decisions for schools, which share the same statistical approaches and many of the same policy decisions.

6.1 Additional measures for teacher reporting

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:

- A multi-year trend is calculated for an individual subject and grade for up to three years.
- A composite is calculated for multiple subjects and grades for up to three years.
- The composite for teachers includes only the subjects for which the teacher has a value-added measure in the current year.
- The composite for teachers weights each subject/grade/year equally.
- The subjects used in the current year are not required to be taught for a consecutive multi-year period.

The composite for teachers will include OAA math and reading. The following examples will be used to show how the OAA multi-year trend in a single subject and the OAA composite across subjects would be calculated for a sample teacher.

6.1.1 Example 1: Available data for OAA multi-year trend for a sample teacher in a single subject

Year	Subject	Grade	Value-Added Gain	Standard Error
2010	Math	8	4.50	1.60
2011	Math	8	3.80	1.50

6.1.2 Example 2: Available data for OAA multi-year composite for a sample teacher across subjects

Year	Subject	Grade	Value-Added Gain	Standard Error
2009	Science	8	4.20	2.00
2009	Math	7	3.50	1.50
2010	Reading	8	0.50	1.40
2010	Math	8	4.50	1.60
2011	Reading	8	-0.30	1.20
2011	Math	8	3.80	1.50

6.2 Calculating gains for the OAA multi-year trend and OAA composite

For the teacher in Example 1, a multi-year trend can be calculated using the two years of data for this teacher in a specific subject and grade, grade eight math. The multiple year trends in OAA math and reading will use re-

estimated value-added gains and standard errors for years prior to the current year. This re-estimation will take into account current year student-level information to provide the most precise and reliable estimate of the prior year using all available information for that teacher in the year being analyzed. Each year used in the OAA multi-year trend is weighted equally, which ensures that teachers are neither advantaged nor disadvantaged due to one particularly different year. Because each group of students and each scenario are different every year, this approach will dampen any year to year variability. Because the value-added estimates are in the same scale (Normal Curve Equivalents), the composite gain across the years is a simple mean gain using all of the cells with equal weights from above.

The multi-year gain for Example 1 is calculated as follows:

$$Multi_{year}Gain = \frac{1}{2}Math_{8_{2010}} + \frac{1}{2}Math_{8_{2011}} = \frac{1}{2}4.50 + \frac{1}{2}3.80 = 4.15 \quad (20)$$

For the teacher in Example 2, a composite gain that includes more than one subject would be calculated. A teacher's composite only includes the subjects for which there is a value-added report in the most recent year. As a result of this policy, the teacher is accountable only for the subject(s) that he or she currently teaches. There are a variety of reasons why a teacher may not teach a particular subject anymore, and this policy mitigates any concerns related to a deliberate decision by the teacher or his/her administrator to focus on other subject(s). As a consequence, this teacher's science report will be excluded from the composite since science had no value-added measure in 2011. Note that science would not be included in a math and reading composite regardless, but this illustrates the point of the subject being there in the current year. However, this teacher's grade seven math report will be included, even though there was no value-added measure for grade seven math in 2011, because there were value-added measures for the subject math in 2011. The last five rows of the chart above represent the five subject/grade/years that will be used in this sample teacher's composite.

Each subject/grade/year used in the OAA composite is weighted equally as was done with the years above. As in Example 1, because the value-added estimates are in the same scale (Normal Curve Equivalents), the composite gain across the five subject/grade/years is a simple mean gain using all of the cells with equal weights. The composite gain is calculated using the following formula:

$$\begin{aligned} Comp\ Gain &= \frac{1}{5}Math_{7_{2009}} + \frac{1}{5}Math_{8_{2010}} + \frac{1}{5}Math_{8_{2011}} + \frac{1}{5}Read_{8_{2010}} + \frac{1}{5}Read_{8_{2011}} \\ &= \frac{1}{5}3.50 + \frac{1}{5}4.50 + \frac{1}{5}3.80 + \frac{1}{5}0.50 - \frac{1}{5}0.30 = 2.40 \end{aligned} \quad (21)$$

6.3 Calculating standard errors for the MRM multi-year trend and OAA composite

The table for each of the two examples above reports a value-added gain estimate as well as a standard error associated with that gain for each subject/grade/year. First of all, please note that 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 multi-year and composite gains that have been calculated, as shown above, from a teacher's value-added gain estimate.

6.3.1 A general formula for the standard error of a composite

First, a bit of terminology: the square of the standard error is called the variance, and it relates to estimation in the context of this document. Statistical formulas are often more conveniently expressed as variances. Standard errors of multi-year trends and 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. Also to maintain generality, the composite is not limited to equally-weighted estimates, but remember that the OAA multi-year trends and composites *do* use equal weighting. Instead each estimate is multiplied by a different weight - a, b, or c.

$$Var(aX + bY + cZ) = a^2Var(X) + b^2Var(Y) + c^2Var(Z) \\ +2ab\ Cov(X,Y) + 2ac\ Cov(X,Z) + 2bc\ Cov(Y,Z) \quad (22)$$

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

$$Cov(X,Y) = Correlation(X,Y)\sqrt{Var(X)}\sqrt{Var(Y)} \quad (23)$$

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

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

6.3.2 Determining statistical independence

Two variables that are unrelated have a correlation, and covariance, of zero. Such variables are said to be statistically independent. This will be the case for multi-year trends as presented in Example 1. If the X and Y values have a positive relationship, then the covariance will also be positive. For the composite gains presented Example 2, the relationship will generally be positive, and this means that the OAA composite standard error is larger than it would be assuming independence.

6.3.3 Example 1: Standard error and index for OAA multi-year value-added gain

As a general rule, two value-added gain estimates are statistically independent if they are based on completely different sets of students. This is almost always the case with multi-year trends in a single subject, which is the case in Example 1. It is unlikely, though not impossible, that any of this teacher's 2010 grade eight math students were also in the teacher's 2011 grade eight math class. With the assumption of independence, the formula for the standard error of the multi-year gain for Example 1 becomes fairly simple. Recall that the standard error is obtained by taking the square root of the variance.

$$Multi_{year}SE\ Gain = \sqrt{\left(\frac{1}{2}\right)^2 \left(SE\ Math_{8_{2010}}\right)^2 + \left(\frac{1}{2}\right)^2 \left(SE\ Math_{8_{2011}}\right)^2} \\ = \frac{1}{2}\sqrt{(1.60)^2 + (1.50)^2} = 1.10 \quad (24)$$

Using the multi-year value-added gain and multi-year standard error, it is possible to calculate an index. The index is simply the value-added gain divided by its standard error. In this example, the index is 4.15 divided by 1.10, which is 3.78 (using the unrounded multi-year standard error).

6.3.4 Example 2: Standard error and index for OAA composite value-added gain

Unlike the situation in Example 1, in this example the standard error of the OAA composite value-added gain cannot be calculated using the assumption that the gains making up the composite are independent. This is

because it is much more likely that some of the same students are represented in different value-added gains, such as grade eight math in 2011 and grade eight reading in 2011. To demonstrate the impact of the covariance terms on the standard, it is useful to calculate the standard error using (inappropriately) the assumption of independence. The standard error would then be as follows:

$$SE\ Comp\ Gain =$$

$$\frac{1}{5}\sqrt{\left(SE\ Math_{7_{2009}}\right)^2 + \left(SE\ Math_{8_{2010}}\right)^2 + \left(SE\ Math_{8_{2011}}\right)^2 + \left(SE\ Read_{8_{2010}}\right)^2 + \left(SE\ Read_{8_{2011}}\right)^2} \quad (25)$$

$$= \frac{1}{5}\sqrt{(1.50)^2 + (1.60)^2 + (1.50)^2 + (1.40)^2 + (1.20)^2} = 0.65$$

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:

$$SE\ Comp\ Gain = \frac{1}{5}\sqrt{\begin{matrix}(1.50)^2 + \\ (1.60)^2 + (1.50)^2 + (1.40)^2 + (1.20)^2 + \\ 2(1.5*1.6 + 1.5*1.5 + 1.5*1.4 + 1.5*1.2 + 1.6*1.5 + \\ 1.6*1.4 + 1.6*1.2 + 1.5*1.4 + 1.5*1.2 + 1.4*1.2)\end{matrix}} = 1.44 \quad (26)$$

The actual standard error will fall somewhere between the two extreme values of 0.65 and 1.44 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 2011 grade eight math and 2010 grade eight math classes had no students in common, then their correlation would be zero. On the other hand, if the 2011 grade eight math and 2011 grade eight 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. 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 of 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 0.65 and 1.44 in this particular 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, or in other words the number of students and the amount of missing data all of those students have will contribute to the magnitude of the standard error.

As in Example 1, the final step is to express the composite value-added gain as an index, calculated by dividing the composite value-added gain by its standard error. In this example, the composite index for this teacher is 2.40 divided by a number between 0.65 and 1.44. If the actual standard error in this example were 0.75, then the index for this teacher would be 2.40 / 0.75 = 3.20. Please note that, while some of the values in the example were rounded for display purposes, the actual rounding or truncating only occurs after all of the measures have been combined, as described in Section 5.3.

6.4 Introduction to school composites and multi-year trends

This section captures how the policy decisions by ODE are implemented in the calculation of composites and multi-year trends for schools in the tested subjects and/or grades. Please note, multi-year trend refers to an average that includes school or teacher value-added estimates from multiple years; however, each school or

teacher value-added estimate is based on up to five years of student data. The decisions for schools share the same statistical approaches and many of the same policy decisions as those for teachers.

The key policy decisions for schools can be summarized as follows:

- A multi-year trend is calculated for an individual subject and grade for up to three years.
- A composite is calculated for multiple subjects and grades for up to three years.
- The composite for schools includes both subjects (OAA math and reading), even if one of those subjects does not have a value-added measure in the current year.
- The composite for schools weights each subject/grade/year equally.

The composite for schools includes OAA math and reading, and because this approach uses both subjects, the statistical approaches are similar to what is described for Example 2. In Example 3 below, the tested subjects and grades for OAA are listed for a sample middle school.

Example 3: Available Data for OAA Multi-Year Composite for a Sample School across Subjects

Year	Subject	Grade	Value-Added Gain	Standard Error
2009	Math	6	4.20	1.00
2009	Reading	6	3.50	0.70
2009	Math	7	2.00	0.90
2009	Reading	7	4.10	0.80
2009	Math	8	1.70	0.90
2009	Reading	8	2.50	0.70
2010	Math	6	-0.50	0.70
2010	Reading	6	4.50	0.80
2010	Math	7	5.00	1.00
2010	Reading	7	2.30	0.80
2010	Math	8	-3.10	0.90
2010	Reading	8	1.20	0.60
2011	Math	6	3.30	0.70
2011	Reading	6	-1.10	1.00
2011	Math	7	2.00	0.50
2011	Reading	7	2.40	1.10
2011	Math	8	-0.30	0.60
2011	Reading	8	3.80	0.70

As in Examples 1 and 2, because the value-added estimates are in the same scale (Normal Curve Equivalents), the school composite gain across the 18 subject/grade/years is a simple mean gain using all of the cells with equal weights. The composite gain is calculated using the following formula:

$$\begin{aligned} Comp\ Gain = {} & \frac{1}{18}Math_{6_{2009}} + \frac{1}{18}Math_{6_{2010}} + \frac{1}{18}Math_{6_{2011}} + \frac{1}{18}Read_{6_{2009}} \\ & + \frac{1}{18}Read_{6_{2010}} + \frac{1}{18}Read_{6_{2011}} + \frac{1}{18}Math_{7_{2009}} + \frac{1}{18}Math_{7_{2010}} \\ & + \frac{1}{18}Math_{7_{2011}} + \frac{1}{18}Read_{7_{2009}} + \frac{1}{18}Read_{7_{2010}} + \frac{1}{18}Read_{7_{2011}} \\ & + \frac{1}{18}Math_{8_{2009}} + \frac{1}{18}Math_{8_{2010}} + \frac{1}{18}Math_{8_{2011}} + \frac{1}{18}Read_{8_{2009}} \\ & + \frac{1}{18}Read_{8_{2010}} + \frac{1}{18}Read_{8_{2011}} = 2.08 \end{aligned} \quad (27)$$

Similar to Example 2, the standard error of the OAA 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 represented in different value-added gains, such as grade eight math in 2011 and grade eight reading in 2011. The statistical approach, outlined in Section 3.1.3 (with references), is quite sophisticated and will take into account the correlations between pairs of value-added gains as shown in equation (22) and using equation (6) for schools and equation (10) for teachers.[2] 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. For the sake of simplicity, let us assume that the actual standard error was 0.40 for the school composite in Example 3.

As in Examples 1 and 2, the final step is to express the school composite value-added gain as an index, calculated by dividing the school composite value-added gain by its standard error. In this example, the composite index for this school is 2.08 divided by 0.40, or 5.20.

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

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 OAA assessments as well as national tests such as college entrance exams ACT for a subset of districts in Ohio. These projections can be used to predict a student's future success (or lack of success) and so may be used to guide counseling and intervention to increase students' likelihood of future success.

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

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i \qquad (28)$$

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

$$\hat{y}_i = \hat{\mu}_y \pm \hat{\alpha}_j + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots + \epsilon_i \qquad (29)$$

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

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

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

$$Prob(\hat{y}_i \geq b) = \Phi\left(\frac{\hat{y}_i - b}{SE(\hat{y}_i)}\right) \quad (30)$$

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 SAS 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 also important that 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 2013, the percentage of students who achieved a maximum score on the OAA assessments ranged from a high of 0.44% (3rd Grade Math) to a low of .015% (6th Grade Reading). 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 of the OAA assessments are well below acceptable values, meaning that the OAA's 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 Ohio Achievement Assessments 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 of these 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 very important when a base year is used to set the expectation of growth since this approach assumes that scale scores mean the same thing in a given subject and grade across years.

Since the Ohio Academic Assessments were not developed specifically to assure the kind of inter-year reliability that is required by value-added, it has been necessary to create a data procedure to accomplish this requirement. In 2009-2010, the scale stabilization procedure was first implemented, and will continue to be implemented for the duration of the use of the OAA's. Although standard psychometric methods are used to provide equivalent scales within a grade and subject, it is recognized that unanticipated variability emerges across grades and among years within a grade in the OAA scaling. Due to the impact this would have on the interpretation of value-added analyses, the scale stabilization procedure was implemented to assure that grade- and subject- level results can be appropriately analyzed and interpreted. The baseline year for the stabilization is 2010 as specified by ODE. Stabilized NCE scores are used in the analyses for grades four through eight in OAA Math and Reading. See Section 3.1.2.2 for more information.

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 that is 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 into 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 1,962,906 records from the 2012-2013 OAA 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 1,962,906 records from the 2012-2013 OAA Math, Reading, and Science assessments, 365 records (0.02%) 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 1,962,906 records from the 2012-2013 OAA Math, Reading, and Science assessments, 37 records (0.002%) 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 of these scores will be excluded from the analysis.

Of the 1,962,906 records from the 2012-2013 OAA Math, Reading, and Science assessments, 4,374 records (0.2%) 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 1,962,906 records from the 2012-2013 OAA Math, Reading, and Science assessments, 114 records (0.006%) 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 2009 to ninth in 2010) or is moved back by one grade or more (i.e. moves from fourth in 2009 to third in 2010) 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, SAS does not remove students with scores that appear to be associated with inconsistent grades. SAS leaves students in the analysis at the tested grade that SAS 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 of 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 math subjects (OAA and OGT) 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, SAS 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 1,962,906 records from the 2012-2013 OAA Math, Reading, and Science assessments, 862 records (0.04%) 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 1,716,331 linkages from the 2012-2013 OAA Math, Reading, and Science assessments, 16,222 linkages (0.95%) 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.