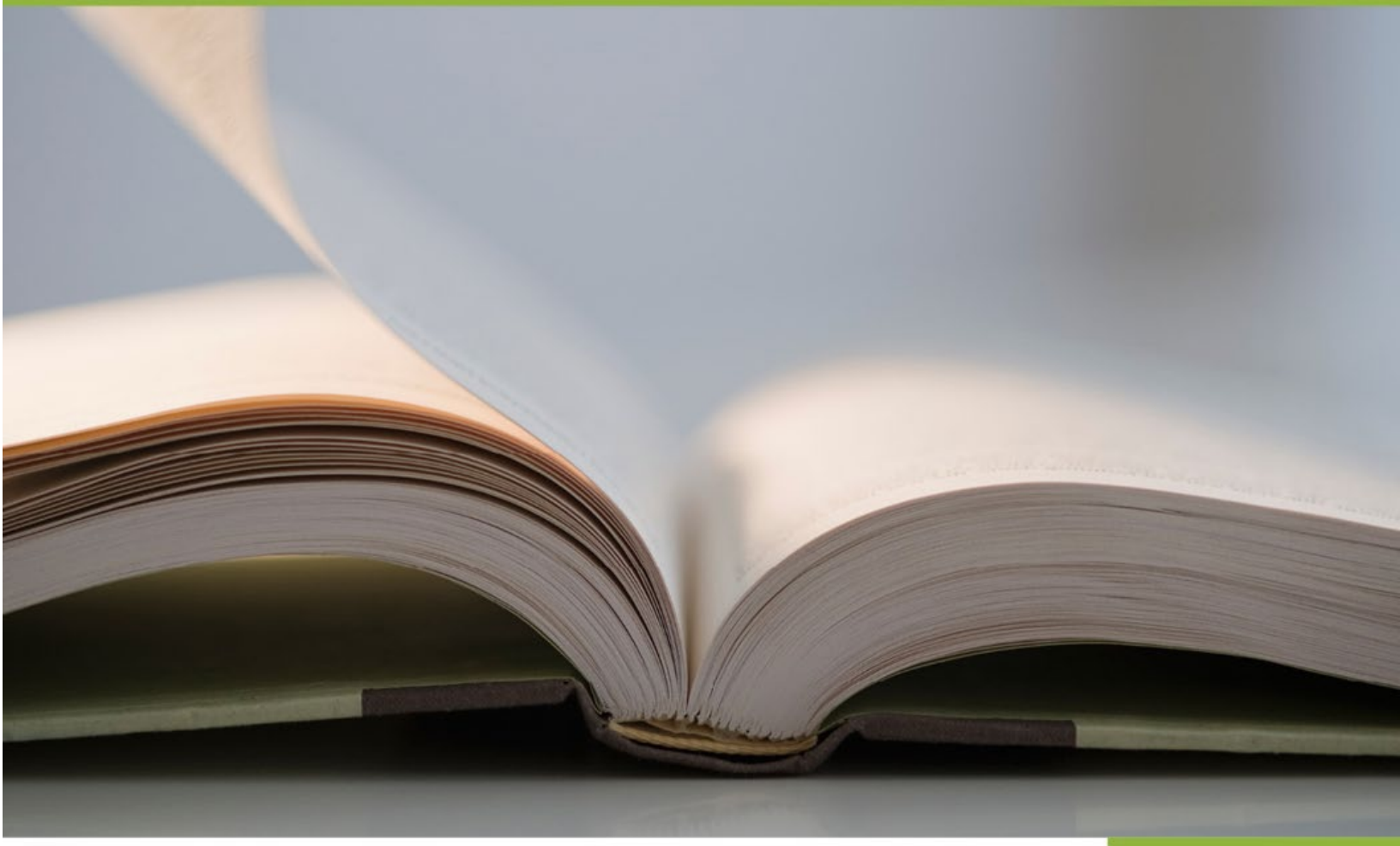
SAS® EVAAS

Statistical Models and Business Rules

Prepared for North Carolina Department of Public Instruction



sas
THE POWER TO KNOW.

Contents

1 Introduction to Value-Added Reporting in North Carolina

Since 2001, EVAAS growth reporting (or value-added reporting) has been available to North Carolina educators and has also been available statewide since 2006. The purpose of EVAAS is to support educators with school improvement through both reflective and proactive planning tools.

Since its inception, EVAAS growth measures focused on the *growth* of students over time rather than their *achievement level*. EVAAS represented a paradigm shift for educators and policymakers and, in identifying the more effective practices and less effective practices, educators receive personalized feedback, which they could then leverage to improve the academic experiences of their students.

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

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

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

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

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

2 Data Inputs

This section provides details about the input data used in the North Carolina value-added model as well as the student, teacher, and school information provided in the assessment files.

2.1 Determining Suitability of Assessments

2.1.1 Current Assessments

To be used appropriately in any value-added analyses, the scales of these tests must meet three criteria. (Additional details about each of these requirements are provided in Section 8.)

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

These criteria are monitored by SAS and psychometricians at NCDPI.

2.2 Assessment Data Used in North Carolina

2.2.1 Assessments

SAS receives the following assessments for EVAAS reporting:

- End-of-grade Math and Reading in grades 3–8
- End-of-grade Science in grades 5 and 8
- End-of-course assessments in Biology, English II, Math 1, and Math 3
- Reading assessments in K–2
- North Carolina Final Exam assessments in various subjects
- Career and Technical Education assessments in various subjects
- ACT assessments in English, Math, Reading, Science, and Composite
- SAT assessments in Evidence-Based Reading and Writing, Math, and Composite
- PSAT assessments in Evidence-Based Reading and Writing and Math
- AP assessments in various subjects

The state End-of-Grade (EOG) tests are administered in the spring semester with the exception of EOG Reading for grade 3, which is tested in both fall and spring. The End-of-Course (EOC) assessments, North Carolina Final Exams (NCFEs), and Career and Technical Education assessments (CTEs) are typically given

in the fall and spring semesters with the occasional summer administration. The K-2 assessments are administered three times throughout the year.

2.2.2 Student Identification Information

SAS receives the following information from NCDPI:

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

2.2.3 Assessment Information Provided

SAS also receives the following information from NCDPI:

- Scale score
- Test taken
- Tested grade
- Tested semester
- District number
- School number
- Membership
  - Accountability Growth Membership
  - Partial Enrollment
- Test Form
- First Year English Learner (EL)

At times, raw scores are provided for the NCFE, and pre-test scores are provided for the CTE assessments.

2.3 Student Information

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

- Academically or Intellectually Gifted (Y, N)
- Gender (M, F)
- English Learners (EL) (Y, 1, 2, U, N)
- Economically Disadvantaged Students (Y, N)
- Students with Disabilities (Y, N)

- Race
  - American Indian/Alaskan Native
  - Asian/Pacific Islander
  - Black (not Hispanic)
  - Hispanic
  - Two or More Races
  - White (not Hispanic)

2.4 Teacher Information

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

Roster verification is completed within the EVAAS web application. NCDPI provides SAS with a file that contains the approved teacher-student linkage data entered into PowerSchool:

- Teacher identification
  - Teacher Name
  - Teacher Unique ID
- Student linking information
  - Student Last Name
  - Student First name
  - Unique Student ID (USID)
- Course information linked to a tested subject via a course to subject mapping provided by DPI
- District and School information (numbers)
- Percentage of instructional responsibility derived from enrollment information provided by DPI (i.e., the date the student enrolled and the date the student left the course)

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 EOG

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 North Carolina, EVAAS includes two main categories of value-added models, each comprised of District, School, and Teacher reports.

- **Multivariate Response Model (MRM**) is used for tests given in consecutive grades, like the EOG Math and Reading in grades 3–8 or the K-2 early grade assessments.
- **Univariate Response Model (URM)** is used for tests given in multiple grades, such as the EOC, NCFE or CTE assessments, or when performance from previous tests is used to predict performance on another test.

Both models offer the following advantages:

- The models include multiple subjects and grades for each student to minimize the influence of measurement error.
- The models can accommodate tests on different scales.
- The models can accommodate students with different sets of testing history.
- The models do not impute any test scores for students who are missing test scores.
- The models can accommodate team teaching or other shared instructional practices.

Each model is described in greater detail in Section 3.1 (MRM) and Section 3.2 (URM) of this document.

Because the EVAAS models use multiple subjects and grades for each student, it is typically 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 and demographic influences persist over time, these influences are already represented in the student's data. As a 2004 study by The Education Trust stated, specifically with regard to the EVAAS modeling:

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

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

The value-added reporting in North Carolina is available for districts, schools, and teachers.

3.1 Multivariate Response Model (MRM)

EVAAS includes 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 it measures growth between two points in time for a group of students. The current growth expectation is met when a cohort of students from grade to grade maintains the same relative position with respect to statewide student achievement in that year for a specific subject and grade. (See Intra-Year Approach in Section 4 for more details.)

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

- All students with valid data are included in the analyses. Each student's testing history is included without imputing any test scores.
- By encompassing all students in the analyses, including those with missing test scores, the model provides the most realistic estimate of achievement available.
- The model minimizes the influence of measurement error inherent in academic assessments by using multiple data points of student test history and multiple years of data.
- The model uses scores from multiple tests, including those on different scales.
- The model accommodates teaching scenarios where more than one teacher has responsibility for a student's learning in a specific subject, grade, and year.
- The model analyzes multiple consecutive grades and 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 studies by experts from RAND Corporation, a non-profit research organization:

- On the **choice of a complex value-added model**: McCaffrey, Daniel F., and J.R. Lockwood. 2008. "Value-Added Models: Analytic Issues." Prepared for the National Research Council and the National Academy of Education, Board on Testing and Accountability Workshop on Value-Added Modeling, Nov. 13-14, 2008, Washington, DC.
- On the **advantages of the longitudinal, mixed model approach**: Lockwood, J.R. and Daniel F. McCaffrey. 2007. "Controlling for Individual Heterogeneity in Longitudinal Models, with Applications to Student Achievement." *Electronic Journal of Statistics* 1: 223-252.
- On the **insufficiency of simple value-added models:** McCaffrey, Daniel F., B. Han, and J.R. Lockwood. 2008. "From Data to Bonuses: A Case Study of the Issues Related to Awarding Teachers Pay on the Basis of the Students' Progress." Presented at Performance Incentives: Their Growing Impact on American K-12 Education, Feb. 28-29, 2008, National Center on Performance Incentives at Vanderbilt University.

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

3.1.1 MRM at the Conceptual Level

An example data set with some description of possible value-added approaches might be helpful for conceptualizing how the MRM works. Assume that 10 students complete a test in two different years with the results shown in Figure 1. The goal is to measure academic growth (gain) from one year to the next. Two simple approaches are to calculate the mean of the differences *or* to calculate the differences of the means. When there is no missing data, these two simple methods provide the same answer (5.80 on the left in Figure 1); however, when there is missing data, each method provides a different result (9.57 versus 3.97 on the right in Figure 1). A more sophisticated model is needed to address this problem.

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

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

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

The MRM uses the correlation between current and previous scores in the nonmissing data to estimate a mean for the set of all previous and all current scores as if there were no missing data. It does this without explicitly assigning values for the missing scores. The difference between these two estimated means is an estimate of the average gain for this group of students. In this small example, the estimated difference on the right is 5.71 when using the MRM approach to first estimate the means in each column and taking the difference. Even in a small example such as this, the estimated difference is much closer to the difference with no missing data (on the left) than either measure obtained by the mean of the differences (3.97) or difference of the means (9.57) on the right. This method of estimation has been shown, on average, to outperform both of the simple methods.[1] In this small example, there were only

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

two grades and one subject. Larger data sets, such as those used in actual EVAAS analyses for North Carolina, provide better correlation estimates by having more student data, subjects, and grades, which in turn provide better estimates of means and gains.

This small example is meant to illustrate the need for a model that will accommodate incomplete data and provide a reliable measure of growth. 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 EVAAS Uses Normal Curve Equivalents in MRM

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

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

For example, in a typical year in North Carolina, the average maximum NCE is approximately 110, corresponding to percentile rankings above 99.0. However, for display purposes in the EVAAS web application and to avoid confusion among users with interpretation, NCEs are shown as integers from 1-99. Truncating would create an artificial ceiling or floor, which might bias the results of the value-added measure for certain types of students. This forces the gain to be close to 0, or even negative, so the actual calculations use non-truncated numbers.

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

By definition, the mean (or average) NCE score for the reference distribution is 50 for each grade and subject. "Growth" is the difference in NCEs from one year/grade to the next in the same subject. The growth standard, which represents a "normal" year's growth, is defined by a value of zero. More specifically, it maintains the same position in the reference distribution from one year/grade to the next**.** It is important to reiterate that a gain of zero on the NCE scale does not indicate "no growth." Rather, it indicates that a group of students in a district, school, or classroom has maintained the same position in the state distribution from one grade to the next. The expectation of growth is set by using each individual year to create NCEs. For more on Growth Expectation, see Section 4.

3.1.2.2 How EVAAS Uses NCEs in MRM

There are multiple ways of creating NCEs. EVAAS MRM uses a method that does not assume that the underlying scale is normal since experience has shown that some testing scales are not normally distributed and this will ensure an equal interval scale. Table 1 provides an example of the way that EVAAS converts scale scores to NCEs.

The first five columns of Table 1 show an example of a tabulated distribution of test scores from North Carolina data. The tabulation shows, for each possible test score, in a particular subject, grade, and year, how many students made that score ("Frequency") and what percentage ("Percent") that frequency was out of the entire student population. (In Table 1, 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 1. If a particular score has a percentile rank of 48, this is interpreted to mean that 48% of students in the population had a lower score and 52% had a higher score. In practice, there is some percentage of students that will receive each specific score. For example, 2.8% of students received a score of 446 in Table 1. The usual convention is to consider half of that 2.8% to be "below" and half "above." Subtracting 1.4% (half of 2.8%) from the 33.6% who scored below the score of 446 produces the percentile rank of 32.2 in Table 1.

Table 1: Converting tabulated test scores to NCE values

Score	Frequency	Cum Freq	Percent	Cum Pct	Ptile Rank	Z	NCE
446	3406	40544	2.8	33.6	32.2	-0.423	40.08
447	5022	45566	4.2	37.8	35.7	-0.312	42.09
448	3589	49155	3.0	40.8	39.2	-0.234	44.07
449	5423	54578	4.5	45.3	43.0	-0.120	46.10
450	3727	58305	3.1	48.3	46.8	-0.042	48.12
451	6037	64342	5.0	53.4	50.9	0.084	50.26
452	4023	68365	3.3	56.7	55.0	0.168	52.47

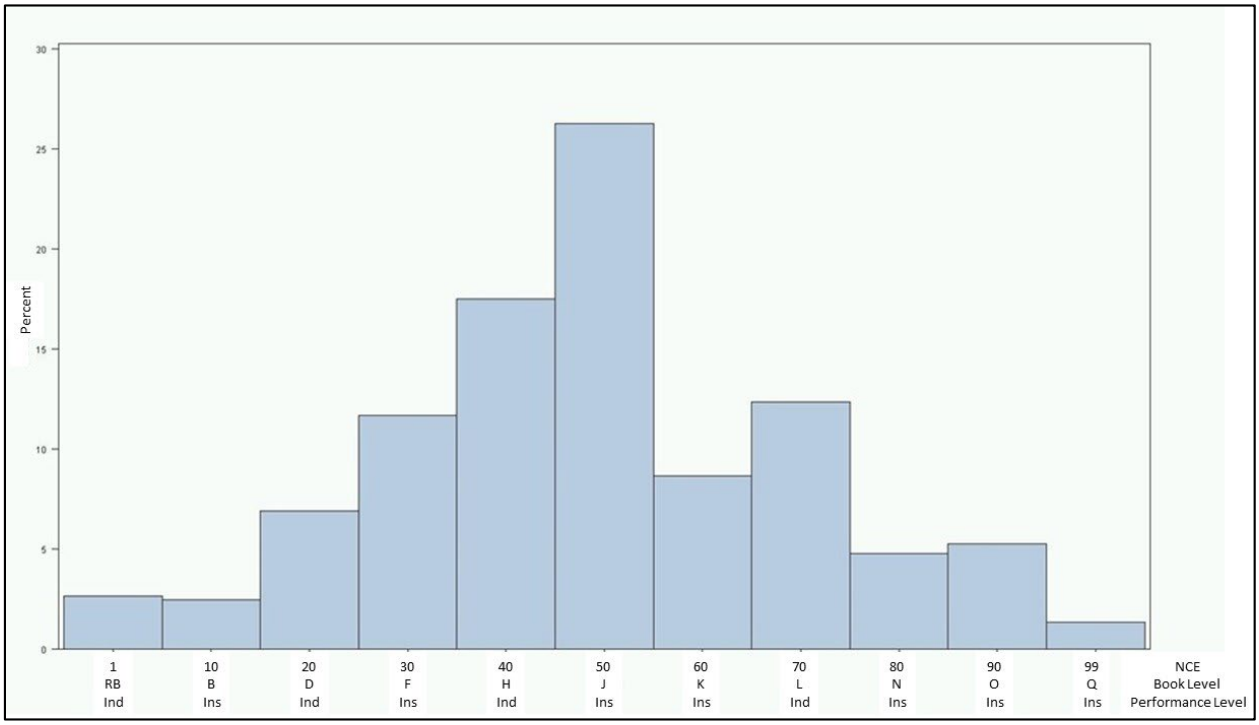
NCEs are obtained from the percentile ranks using the normal distribution. Using a table of the standard normal distribution (found in many textbooks) or computer software (for example, a spreadsheet), one can obtain the associated Z-score from a standard normal distribution for any given percentile rank. 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). NCEs are further adjusted by considering a statewide MRM model and accounting for missing test scores to ensure that the average achievement on the NCE scale is 50 for each subject and grade modeled.

3.1.2.3 How EVAAS Uses NCEs in the K-2 Assessment

NCEs can also be created for assessments where the underlying scale is not inherently numeric in nature. One such assessment is the K-2 Text Reading and Comprehension assessment, which presents student achievement results in book levels and performance levels. Book levels range from Print Concepts (PC), Reading Behaviors (RB), B, C and so on up to U. PC is the lowest possible book level, and U is the highest possible book level on the distribution of possible book levels. Furthermore, each book level has three performance levels corresponding to the student's reading and comprehension mastery of the text: Frustrational, Instructional, and Independent. Even though book levels and performance levels are non-numeric, the combination of the two provides the measured reading and comprehension ability of the test taker.

The frequencies of all observed book levels and performance levels of a population of test takers can be aggregated in an overall scoring distribution where each book and performance level are translated to corresponding percentiles and NCEs just as the case with other assessments that report numeric scale scores. NCEs for the K-2 Assessment in North Carolina are calculated by grade and benchmark period: Beginning-of-Year (BOY), Middle-of-Year (MOY), and End-of-Year (EOY). For example, Figure 2 displays the NCEs associated with book and performance levels and the frequency of each level for the 2018 Grade 2 EOY Text Reading and Comprehension assessment.

Figure 2: NCEs for the 2018 2nd Grade EOY Text Reading and Comprehension assessment



Growth for the K-2 Assessment is the difference in NCEs from a starting benchmark period (MOY for Kindergartners, BOY for first and second grade) to the EOY benchmark period. The average NCE for a district, school or classroom can be compared to the overall amount of growth exhibited in the state, which represents a "normal" year's growth, otherwise known as the growth standard.

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 (NCEs) for all students in multiple academic subjects tested over all grades and years.

X is a known $m \times p$ matrix that allows the inclusion of any fixed effects. Fixed effects are factors within the model that come from a finite population, such as all of the individual schools in the state of North Carolina. In the school model, there is a fixed effect for every school/year/subject/grade. This matrix would have a row for each of these combinations.

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

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

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

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

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

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

where R is the $m \times m$ matrix that reflects the correlation among the student scores residual to the specific model being fitted to the data, and G is the $q \times q$ variance-covariance matrix that reflects the correlation among the random effects. If (v, ϵ) are normally distributed, the joint density of (y, v) is maximized when β has value b and v has value u given by the solution to the following equations, known as Henderson's mixed model equations:[2]

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

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

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

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

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.

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

3.1.3.1 District and School Models

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)

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

with a column representing each interaction of school (in the school model), subject, grade, and year of data. The fixed-effects vector β contains the mean score for each school, subject, grade, and year, with each element of β corresponding to a column of X. Since MRMs are generally run with each school uniquely defined across districts, there is no need to include district in the model.

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

Algebraically, the district MRM is represented as:

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

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

The school MRM is represented as:

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

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

The MRM uses multiple years of data to estimate the covariances that can be found in the matrix R_0. This estimation of covariances is done within each level of analyses and can result in slightly different values within each analysis. Each level of analysis will use the values found within that analysis.

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

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

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

3.1.3.2 Teacher Model

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

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

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

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

Algebraically, the teacher model is represented as:

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

y_{ijkl} is the test score for the i^{th} student in the j^{th} subject in the k^{th}grade in the l^{th} year. $\tau_{ijk^*l^*t}$ is the teacher effect of the t^{th} teacher on the i^{th} student in the j^{th} subject in grade k^* in year l^*. The complexity of the parenthetical term containing the teacher effects is due to two factors. First, in any given subject/grade/year, a student might have more than one teacher. The inner (rightmost) summation is over all the teachers of the i^{th} student in a particular subject/grade/year. $\tau_{ijk^*l^*t}$ is the

effect of those teachers. $w_{ijk^*l^*t}$ is the fraction of the i^{th} student's instructional time claimed by the t^{th} teacher. Second, as mentioned above, this model allows teacher effects to accumulate over time. That is, how well a student does in the current subject/grade/year depends not only on the current teacher but also on the accumulated knowledge and skills acquired under previous teachers. The outer (leftmost) summation accumulates teacher effects not only for the current (subscripts k and l) but also over previous grades and years (subscripts k^* and l^*) in the same subject. Because of this accumulation of teacher effects, this type of model is often called the "layered" model.

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

Table 2 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).

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

Susan is taught by two different teachers in grade 3, teacher Abbot for Math and, teacher Banks for Reading. In grade 4, Susan had teacher Card for reading. For some reason, in grade 4 no teacher claimed Susan for Math even though Susan had a grade 4 Math test score. This score can still be included in the analysis by entering zeros into the Susan's Z matrix rows for grade 4 Math. In grade 5, on the other hand, Susan had no test score in Reading. This row is completely omitted from the Z matrix. There will always be a Z matrix row corresponding to each test score in the y vector. Since Susan has no entry in y for grade 5 Reading, there can be no corresponding row in Z.

Eric's scenario illustrates team teaching. In grade 3 Reading, Eric received an equal amount of instruction from both teachers Abbot and Banks. The entries in the Z matrix indicate each teacher's contribution, 0.5 for each teacher. In grade 5 Math, however, while Eric was taught by both teachers East and Farr, they did not make an equal contribution. Teacher East claimed 80% responsibility and teacher Farr claimed 20%.

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

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

The teacher model provides estimated mean gains for each subject and grade. These quantities can be described by linear combinations of the fixed and random effects and are found using the equations mentioned above.

Table 2: Encoding the Z matrix

			Teachers											
			Third Grade				Fourth Grade				Fifth Grade			
			Abbot		Banks		Card		Dupont		East		Farr	
Student	Grade	Subjects	M	R	M	R	M	R	M	R	M	R	M	R
Tommy	**3**	**M**	1	0	0	0	0	0	0	0	0	0	0	0
		R	0	1	0	0	0	0	0	0	0	0	0	0
	4	**M**	1	0	0	0	1	0	0	0	0	0	0	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 North Carolina

The MRM is used with the EOG test in Math and in Reading for grades 3–8 to provide value-added measures for districts, schools, and teachers in grades 4–8 in Math and grades 3-8 in Reading. The MRM is also used with the K-2 assessment in Reading for K–2 to provide value-added measures for districts, schools, and teachers in those grades.

The MRM methodology provides estimated measures of growth for up to three years in each subject/grade/year for district, school, and teacher analyses provided that the minimum student requirements are met. (Details are in Section 3.1.6.) For each subject, measures are also given across grades, across years (up to three-year averages), and combined across grades and years.

For teachers, value-added measures for each EOG or K-2 subject/grade/year are computed (and displayed on the EVAAS web application available at https://ncdpi.sas.com/).

More information about teacher composite measures can be found in Section 6.

3.1.5 Students Included in the Analysis

All students' scores are included in these analyses if the scores can be used and do not meet any criteria for exclusion outlined in Section 8. In other words, a complete history of every student's Math and Reading results for the student's cohort are incorporated into the models.

There are some exclusion rules based on policy decisions by NCDPI. For the MRM, student scores are excluded from the analyses if the student is flagged as a First Year EL student, and students must meet partial enrollment membership to be included in the analysis.

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

3.1.5.1 District and School Measures

3.1.5.1.1 Overall Measures of Student Growth for Districts and Schools

The analyses for schools and districts include all applicable student scores from EOG math and reading tests from the cohort of students testing in the most recent three years or all applicable student scores from K-2 for early grade reporting.

3.1.5.1.2 Student Group Measures of Student Growth for Districts and Schools

Student group value-added measures are included in North Carolina's federal accountability system. This includes the following student groups:

- American Indian/Alaskan Native
- Asian/Pacific Islander
- Black (not Hispanic)
- Hispanic

- Two or More Races
- White (not Hispanic)
- Economically Disadvantaged Students (EDS)
- English Learners (EL)
- Students with Disabilities (SWD)
- Academically or Intellectually Gifted (AIG)

Students are identified as members of a group based on a flag in the student record. Growth measures are calculated for each subset of students for each district and school that meet the minimum requirements of student data.

In each student group value-added computation, the expectation of growth is defined the same as in the overall students' analysis. In other words, the expectation of growth is based on all students. Furthermore, the estimated covariance parameters are used from the overall students' analysis when calculating the value-added measures. These measures are provided using the EOG subjects with a composite across Math in grades 4–8 and Reading in grades 4–8.

3.1.5.2 Teacher Measures

The Teacher Value-Added reports use all available test scores for each individual student linked to a teacher through the roster verification process unless a student or a student's test score meets certain criteria for exclusion.

3.1.6 Minimum Number of Students for Reporting

3.1.6.1 District and School Models

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

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

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

3.1.6.2 Teacher Model

The teacher value-added model includes teachers who are linked to at least six students with a valid test score in the same subject and grade. To clarify, this means that the teachers are included in the analysis, even if they do not receive a report due to the other requirements. This requirement does not consider the percentage of instructional time the teacher spends with each student in a specific subject/grade.

However, to receive a teacher value-added *report* for a particular year, subject, and grade, there are two additional requirements. First, a teacher must have at least six Full Time Equivalent (FTE) students in a specific subject/grade/year. The teacher's number of FTE students is based on the number of students linked to that teacher and the percentage of instructional time the teacher has for each student. For example, if a teacher taught 10 students for 50% of their instructional time, then the teacher's FTE number of students would be five, and the teacher would not receive a Teacher Value-Added report. If another teacher taught 12 students for 50% of their instructional time, then that teacher would have six FTE students and would receive a Teacher Value-Added report. The instructional time attribution is obtained from the student-teacher linkage data. This information is in the files sent to EVAAS described in Section 2.

As the second requirement, the teacher must be linked to at least five students with prior test score data in the same subject, and the test data might come from any prior grade as long as they are part of the student's regular cohort. (If a student repeats a grade, then the prior test data would not apply as the student has started a new cohort.) One of these five students must have a "simple gain," meaning the same subject prior test score must come from the immediate prior year and prior grade. Students are linked to a teacher based on the subject area taught and the assessment taken.

3.1.7 Hurricane Florence Adjustment

At the request of NCDPI, SAS conducted an analysis to assess whether students' growth measures were related to their districts' loss of instructional days due to Hurricane Florence in the 2018-19 school year. This analysis indicated a need to adjust the growth model for EOG Reading in grade 3 to ensure validity and comparability of results statewide. As a result, the growth model for EOG Reading in grade 3 makes an adjustment to students' Beginning-of-Year (BOY) test score based on the number of days missed and waived due to the hurricane as well as students' performance on other assessments, such as their prior test scores in grade 2 and their End-of-Year (EOY) test score in grade 3. In technical terms, the growth model uses linear regression to establish a relationship among grade 2 test scores, grade 3 test scores, and the number of days missed and waived. The BOY test scores are adjusted prior to use in the growth model.

3.2 Univariate Response Model (URM)

Tests that are not necessarily administered to students in consecutive years, like the EOC and CTE tests, require a different modeling approach from the MRM, and this modeling approach is called the univariate response model (URM) or predictive model. This model is also used when previous test performance is used to predict another test's performance, such as the NCFE or ACT. The statistical model can also be classified as a linear mixed model and can be further described as an analysis of covariance (ANCOVA) model. The URM is a regression-based model, which measures the difference between students' predicted scores for a particular subject/year with their observed scores. The growth expectation is met when students with a district/school/teacher made the same amount of growth 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 compare to the average of those teachers with students taking that assessment, such as the case with many CTE assessments and some NCFE assessments.

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

- The model does not require students to have all predictors or the same set of predictors as long as a student has at least three prior test scores in any subject/grade.

- The model minimizes the influence of measurement error by using many prior tests for an individual student. Analyzing all subjects simultaneously increases the precision of the estimates.
- The model uses scores from multiple tests, including those on different scales.
- The model accommodates teaching scenarios where more than one teacher has responsibility for a student's learning in a specific subject/grade/year.

In North Carolina, URM value-added reporting is available for NCFE, CTE, ACT, SAT, PSAT, and all EOC assessments for districts and schools. Teacher measures are also available for EOC, NCFE, and CTE assessments.

3.2.1 URM at the Conceptual Level

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

Once that relationship has been defined, a predicted score can be calculated for each individual student based on his or her own prior testing history. With each predicted score based on a student's prior testing history, this information can be aggregated to districts, schools, or teachers. 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 growth, so that if every student obtained their predicted score, a district, school, or teacher would likely receive a value-added measure close to zero. A negative or zero value does not mean "zero growth" since this is all relative to what was observed in the state (or pool) that year.

3.2.2 Technical Description of the District, School, and Teacher Models

The URM has similar models for district and school and a slightly different model for teachers that allows multiple teachers to share instructional responsibility. The approach is described briefly below, with more details following.

- The score to be predicted serves as the response variable (y, the dependent variable).
- The covariates (xs, predictor variables, explanatory variables, independent variables) are scores on tests the student has already taken.
- The categorical variable (class variable, factor) are the teacher(s) from whom the student received instruction in the subject/grade/year of the response variable (y).

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

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

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

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

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

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

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

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

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

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

Once the parameter estimates for the prediction equation have been obtained, predictions can be made for any student with any set of predictor values as 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 γs.

3.2.3 Students Included in the Analysis

3.2.3.1.1 Overall Measures of Student Growth for Districts, Schools, and Teachers

In order for a student's score to be used in the district or school analysis for a particular subject/grade/year, the student must have at least three valid predictor scores that can be used in the analysis, all of which cannot be deemed outliers. These scores can be from any year, subject, and grade 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 the student is excluded from the analyses. It is important to note 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.

There are some exclusion rules based on policy decisions by NCDPI. For the URM, student scores are excluded from the analyses if the student is flagged as a First Year EL student or if the student does not meet partial enrollment membership for EOC, NCFE and CTE assessments. For the Math 3 value-added reporting, there are two sets of school reports: one set that excludes students as described for EOCs and another set that further excludes students based on a School Accountability Growth flag for EOC Math 3. This flag indicates whether the student was previously used in School Accountability Growth for Math 1 and should therefore be excluded from School Accountability Growth for Math 3. Note that Teacher reports based on Math 3 do not exclude students based on the School Accountability Growth flag. There are no membership rules used to include or exclude students in the SAT, PSAT, and ACT analyses.

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

3.2.3.1.2 Student Group Measures of Student Growth for Districts and Schools

Student group value-added measures are included in North Carolina's federal accountability system. This includes the following student groups:

- American Indian/Alaskan Native
- Asian/Pacific Islander
- Black (not Hispanic)
- Hispanic
- Two or More Races
- White (not Hispanic)
- Economically Disadvantaged Students (EDS)
- English Learners (EL)
- Students with Disabilities (SWD)
- Academically or Intellectually Gifted (AIG)

Students are identified as members of a group based on a flag in the student record. Growth measures are calculated for each subset of students for each district and school that meet the minimum requirements of student data.

In each student group value-added computation, the expectation of growth is defined the same as in the overall students' analysis. In other words, the expectation of growth is based on all students. Furthermore, the estimated covariance parameters are used from the overall students' analysis when calculating the value-added measures. These measures are provided using the EOC subjects with a composite across Math 1, Math 3, and English II. The Math 3 student group reporting includes only students who meet the accountability business rules described in the second set of reports described in Section 3.2.3.1.1.

3.2.4 Minimum Number of Students for Reporting

To receive an overall measure of student growth, a district or school must have at least 10 students in that year, subject, and grade that have the required three prior test scores needed to obtain a predicted score in that year, subject, and grade and have met all other requirements to be included. Student group reporting also requires 10 students to be included in the EVAAS web reporting.

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

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 EOG tests

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

However, the precise definition of "expected growth" varies by model, and this section provides more detail.

4.1 Intra-Year Growth Expectation

4.1.1 Description

- The actual definitions in each model are slightly different, but the concept can be considered as the average amount of growth seen across the state in a statewide implementation.
- Using the URM model, the definition of the expectation is that students with a district, school, or teacher made the same amount of growth as students with the average district, school, or teacher in the state for that same year/subject/grade. If not all students are taking an assessment in the state, then it might be a subset.
- Using the MRM model, the definition of this type of expectation of growth is that students maintained the same relative position with respect to the statewide student achievement from one year to the next in the same subject area. For example, if students' achievement was at the 50th NCE in 2018 grade 4 Math, based on the 2018 grade 4 Math statewide distribution of student achievement, and their achievement is at the 50th NCE in 2019 grade 5 Math, based on the 2019 grade 5 Math statewide distribution of student achievement, then their estimated gain is 0.0 NCEs.
- With this approach, the value-added measures tend to be centered on the growth expectation every year, with approximately half of the district/school/teacher estimates above zero and approximately half of the district/school/teacher estimates below zero. However, it should be noted that there is not a set distribution of the value-added measures. Being centered on the growth expectation does not mean half of the measures would be in the positive levels and half would be in the negative levels since many value-added measures are indistinguishable from the expectation when considering the statistical certainly around that measure. More details can be found in Section 5.

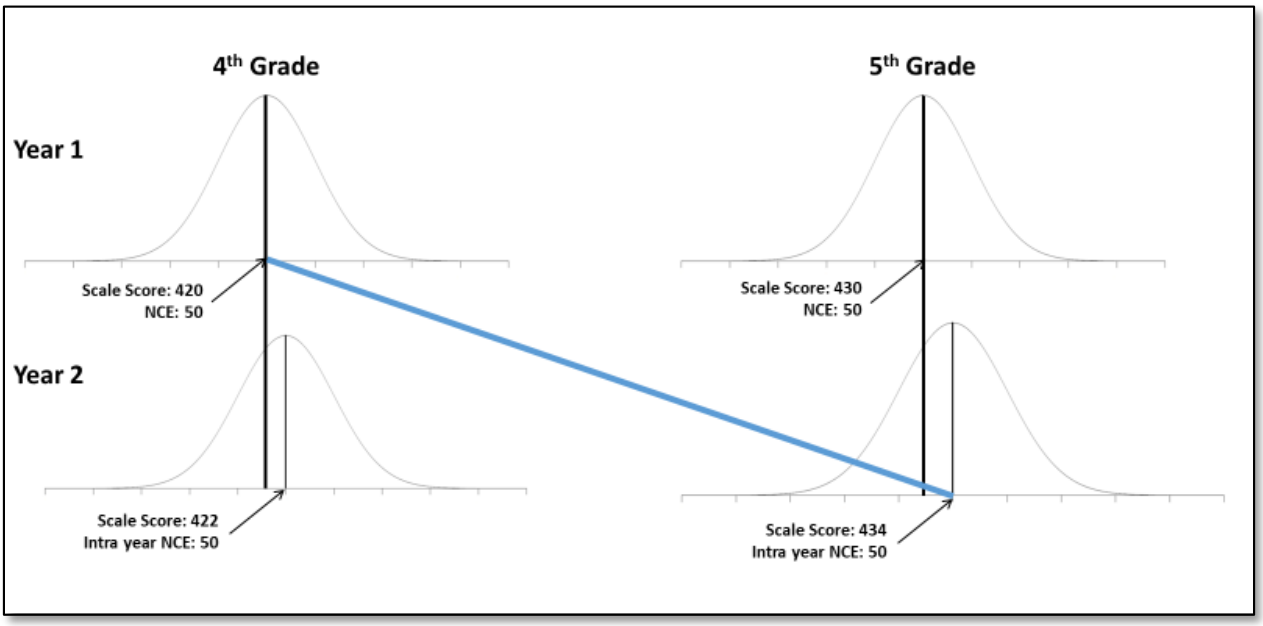
4.1.2 Illustrated Example

Figure 3 below provides a simplified example of how growth is calculated with an intra-year approach when the state achievement increases. The figure has four graphs, each of which plot the NCE distribution of scale scores for a given year and grade. The scale scores are used to illustrate an example in the graphics below and do not represent actual scale scores in North Carolina. In this example, the figure shows how the gain is calculated for a group of grade 4 students in Year 1 as they become grade 5 students in Year 2. In Year 1, our grade 4 students score, on average, 420 scale score points on the test, which corresponds to the 50th NCE (similar to the 50th percentile). In Year 2, the students score, on

average, 434 scale score points on the test, which corresponds to a 50th NCE *based on the grade 5 distribution of scores in Year 2*. The grade 5 distribution of scale scores in Year 2 was higher than the grade 5 distribution of scale scores in Year 1, which is why the lower right-hand graph is shifted slightly to the right. The blue line shows what is required for students to make expected growth, which would be to maintain their position at the 50th NCE in grade 4 in Year 1 as they become grade 5 students in Year 2. The growth measure for these students is Year 2 NCE – Year 1 NCE, which would be 50 – 50 = 0. Similarly, if a group of students started at the 35th NCE, the expectation is that they would maintain that 35th NCE.

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

Figure 3: Intra-year approach example



4.2 Defining the Expectation of Growth During an Assessment Change

During the change of assessments, the scales from one year to the next will be completely different from one another. This does not present any particular changes with the URM methodology because all predictors in this approach are already on different scales from the response variable, so the transition is no different from a scaling perspective. Of course, there will be a need for the predictors to be adequately related to the response variable of the new assessment, but that typically is not an issue.

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

Over the past 20 years, EVAAS reporting has accommodated several different changes in testing regimes and used several tests for the MRM without a break in reporting, such as the change in assessments in North Carolina in 2012.

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

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

5.1 Using Standard Errors Derived from the Models

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

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

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

5.2 Defining Effectiveness in Terms of Standard Errors

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

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

The chart below provides the color-coding, definitions, and interpretation for the Value-Added reports for teachers, which are similar to those provided for districts and schools.

Value-Added Color and Teacher Measure Designation	Growth Measure Compared to the Growth Standard	Index*	Interpretation
Exceeds Expected Growth	At least 2 standard errors above	2.00 or greater	Significant evidence that students made more progress than the Growth Standard.
Meets Expected Growth	Between 2 standard errors above and 2 standard errors below	Between -2.00 and 2.00	Evidence that students made progress similar to the Growth Standard.
Does Not Meet Expected Growth	More than 2 standard errors below	Less than -2.00	Significant evidence that students made less progress than the Growth Standard.

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

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

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

5.3 Rounding and Truncating Rules

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

When value-added measures are also 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 EVAAS Composite Calculations

6.1 Introduction

This section describes how the policy decisions by NCDPI are implemented in the calculation of composites for teachers and schools in the tested subjects and/or grades.

The key policy decisions for teacher composites can be summarized as follows:

- This composite is called the Student Growth Measure, and it includes all available growth measures associated with teachers who received value-added reports within the past three consecutive reporting years.
- For each reporting year, a single-year composite is calculated by weighing each subject/grade/year (for EOG, K-2 and NCFE) and each subject/year (for EOC and CTE) according to the effective number of students' scores included in the value-added measure.
- The composite is then a simple average with equal weighting given to each single-year composite.

The key policy decisions for school composites can be summarized as follows:

- A composite is calculated across subjects and grades using one year of growth measures for schools.
- There are two types of school composites. The first is the School Accountability Growth (SAG) composite that includes only EOC and EOG subjects and grades. Biology is not included in SAG. There is a second composite, Educator Effective Growth (EEG), which uses more subjects and grades associated with a school since it also includes K-2, NCFE, and CTEs.
- Both the SAG and EEG composites weigh each subject/grade/year (for EOG, K-2 and NCFE) and each subject/year (for EOC and CTE) according to the number of scores included in the value-added measures.

A composite combines value-added measures from different tests, subjects, and grades. The following sections show how a Student Growth Measure composite is calculated for a sample teacher. Although we present a teacher example, the process for school composite calculations is the same.

6.2 Teacher Composites

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

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

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

Table 3: Sample teacher value-added information

Year	Subject	Grade	Value-Added Measure	Standard Error	Number of FTE Students
2019	EOG Reading	8	-0.30	1.20	65
2019	EOG Math	8	3.80	1.50	70
2019	Math 1	8	11.75	6.20	20

6.2.1 Calculate MRM-Based Composite Gain Across Subjects

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

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

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

6.2.2 Calculate MRM-Based Standard Error Across Subjects

6.2.2.1 Technical Background on Standard Errors

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

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

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

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

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

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

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

Two variables that are unrelated have a correlation, and covariance, of zero. Such variables are said to be statistically independent. If the X and Y values have a positive relationship, then the covariance will also be positive. As a general rule, two value-added gain estimates are statistically independent if they are based on completely different sets of students. For our sample teacher's MRM composite gain, the relationship will generally be positive, and this means that the MRM-based composite standard error is larger than it would be assuming independence.

6.2.2.2 Illustration of MRM-Based Standard Error for a Sample Teacher

For the sample teacher, it cannot be assumed that the gains in the composite are independent because it is likely that some of the same students are represented in different value-added gains, such as grade 8 Math in 2019 and grade 8 Reading in 2019.

However, to demonstrate the impact of the covariance terms on the standard error, it is useful to calculate the standard error using (inappropriately) the assumption of independence. Using the MRM-based FtE weightings and standard errors reported in Table 3 and assuming total independence, the standard error would then be as follows:

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

At the other extreme, if the correlation between each pair of value-added gains had its maximum value of +1, the standard error would be as follows using the MRM-based FtE weightings and standard errors from Table 3:

$$MRM\ Comp\ SE = \sqrt{\left(\frac{65}{135}\right)^2 (SE\ Read_8)^2 + \left(\frac{70}{135}\right)^2 (SE\ Math_8)^2 + 2\left(\frac{65}{135}\right)\left(\frac{70}{135}\right)(SE\ Read_8)(SE\ Math_8)} \quad (24)$$

$$= \sqrt{\left(\frac{65}{135}\right)^2 (1.20)^2 + \left(\frac{70}{135}\right)^2 (1.50)^2 + 2\left(\frac{65}{135}\right)\left(\frac{70}{135}\right)(1.20)(1.50)} = 1.36$$

The actual standard error will fall somewhere between the two extreme values of 0.97 and 1.36 with the specific value depending on the values of the correlations between pairs of value-added gains. The magnitude of each correlation depends on the extent to which the same students are in both estimates for any two subject/grade/year estimates. For example, if the 2019 grade 8 Math and 2019 grade 8 Reading classes had no students in common, then their correlation would be zero. If the 2019 grade 8 Math and 2019 grade 8 Reading classes contained many of the same students, there would be a positive correlation. However, even if those two classes had exactly the same students, the correlation would likely be considerably less than +1. 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.

Although 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 is typically between these two extremes, which are 0.97 and 1.36 in this particular example. In general, the standard error of the composite gain will vary depending on the standard errors of the value-added gains and the correlations between pairs of value-added gains. The standard errors of the individual value-added gains will depend on the quantity and quality of the data that went into the gain, such as the number of students and the amount of missing data all of those students have, will contribute to the magnitude of the standard error.

6.2.3 Calculate MRM-Based Composite Index Across Subjects

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

$$MRM\ Comp\ Index = \frac{MRM\ Comp\ Gain}{MRM\ Comp\ SE} = \frac{1.83}{1.15} = 1.59 \quad (25)$$

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

6.2.4 Calculate URM-Based Index Across Subjects

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

$$URM\ Comp\ Index = \frac{Math\ 1\ VA\ Measure}{Math\ 1\ SE} = \frac{11.75}{6.20} = 1.90 \quad (26)$$

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

6.2.5 Calculate the Combined MRM and URM Composite Index Across Subjects

The two composite indices from the MRM and URM are then weighted according to the number of students linked to each model to determine the combined composite index. Our sample teacher has 155 students, of which 135 are linked to the MRM and 20 to the URM, so the combined composite index would be calculated as follows using these weightings, the MRM-based composite index across subjects, and the URM-based index across subjects:

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

This combined index is not an actual index itself until it is adjusted to accommodate for the fact that it is based on multiple pieces of evidence together. An index by definition has a standard error of 1, but this unadjusted value (1.62) does not have a standard error of 1. The next step is to calculate the new standard error and divide the combined composite index found above by it. This new, adjusted composite index will be the final index with a standard error of 1. The standard error can be found given the standard formula above and the fact that each index has a standard error of 1. Independence is assumed since these are done outside of the models. In this example, the standard error would be as follows:

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

Therefore, the final combined composite index value is 1.62 divided by 0.88 or 1.85. This is the value in the teacher's SGM report. If this teacher had three consecutive years of growth measures, then each yearly composite is estimated by the process outlined above, and the teacher's SGM is a simple average of the three single-year composites.

6.3 School Composites

The composites calculated for schools are done in the exact same way as teachers described in the section above based on the applicable growth measures.

7 EVAAS Projection Model

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

Currently, the following projections are available to educators in North Carolina:

- EOG Reading in grades 3-8
- EOG Math grades 4–8
- EOG Science in grades 5 and 8
- EOC Math 1, Math 3, Biology I, and English II
- ACT Composite, English, Math, Reading, and Science
- SAT Composite, Evidence-Based Reading and Writing, and Math
- PSAT Composite, Evidence-Based Reading and Writing, and Math
- CTE in various subjects
- NCFE in various subjects
- AP in various subjects

Projections are made one or two grades above the last tested grade for EOG Reading and Math and to the next tested subject/grade or course for EOG Science, EOC, CTE, NCFE, ACT, SAT, PSAT, and AP.

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

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i \quad (29)$$

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

$$\hat{y}_i = \hat{\mu}_y \pm \hat{\alpha}_j + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots + \epsilon_i \quad (30)$$

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. Therefore, this term is usually omitted from the projections. This is

equivalent to setting $\hat{\alpha}_j$ to zero, that is, to assuming that the student encounters the "average schooling experience" in the future.

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 level 3 on a future EOG test, or the probability of scoring sufficiently well on a college entrance exam to gain admittance into a desired program.

Projections are made to levels 2–5 for the EOG and EOC tests, to the proficient level on the CTE tests only for students enrolled in those courses, the 50th and 80th percentile for the NCFE assessments, and to a level of 3 or higher, 4 or higher, or 5 on the AP assessments. Using college readiness assessments, projections are made to US and state averages for PSAT, ACT, and SAT and to the average ACT and SAT scores for incoming NC State University freshmen at various NCSU colleges.

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.

$$Prob(\hat{y}_i \geq b) = \Phi\left(\frac{\hat{y}_i - b}{SE(\hat{y}_i)}\right) \quad (31)$$

8 Data Quality and Pre-Analytic Data Processing

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

8.1 Data Quality

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

8.2 Checks of Scaled Score Distributions

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

8.2.1 Stretch

Stretch indicates whether the scaling of the test permits student growth to be measured for either very low- or very high-achieving students. A test "ceiling" or "floor" inhibits the ability to assess growth for students who would have otherwise scored higher or lower than the test allowed. There must be enough test scores at the high or low end of achievement for measurable differences to be observed. Stretch can be determined by the percentage of students who score near the minimum or the maximum level for each assessment. If a large percentage of students scored at the maximum in one grade compared to the prior grade, then it might seem that these students had negative growth at the very top of the scale. However, this is likely due to the artificial ceiling of the assessment. Percentages for all North Carolina state assessments ultimately used in calculating growth measures are suitable for value-added analysis; this means that the state tests have adequate stretch to measure value-added even in situations where the group of students are very high or low achieving.

8.2.2 Relevance

Relevance indicates whether the test has sufficient alignment with the state standards. The requirement that tested material will correlate with standards if the assessments are designed to assess what students are expected to know and be able to do at each grade level. This is how state tests are designed and is monitored by NCDPI and their psychometricians.

8.2.3 Reliability

Reliability can be viewed in a few different ways for assessments. Psychometricians view reliability as the idea that a student would receive similar scores if they took the assessment multiple times. This type of reliability is important for most any use of standardized assessments.

8.3 Data Quality Business Rules

More information about pre-analytic processing for student test scores is detailed below.

8.3.1 Missing Grade Levels

In North Carolina, 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 K-2 or EOG 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.

8.3.2 Duplicate (Same) Scores

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

8.3.3 Students with Missing Districts or Schools for Some Scores but Not Others

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

8.3.4 Students with Multiple (Different) Scores in the Same Testing Administration

If a student has multiple scores in the same period for a particular subject and grade and the test scores are not the same, then those scores will be excluded from the analysis. If duplicate scores for a particular subject and tested grade in a given testing period are at different schools, then both of these scores will be excluded from the analysis. The highest composite combination of SAT subjects is used for SAT value-added and student college readiness projections.

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 student data are adjusted so that each unique student is associated with only the appropriate scores. If the scores appear to all be associated with a single unique student, then scores that appear inconsistent are excluded from the analysis.

8.3.6 Students with Records That Have Unexpected Grade Level Changes

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

8.3.7 Students with Records at Multiple Schools in the Same Test Period

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

8.3.8 Outliers

8.3.8.1 Conceptual Explanation

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

8.3.8.2 Technical Explanation

Student assessment scores are checked each year to determine whether they are outliers in context with 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, Math subjects (EOG and EOC assessments) are examined simultaneously during outlier identification for the state assessments, and any scores that appear inconsistent, given the other scores for the student, are flagged. Outlier identification for college readiness assessments use all available college readiness data alongside state assessments in the respective subject area (e.g., Math subjects with EOC, EOG, and PSAT tests might be used to identify outliers with SAT or ACT). Furthermore, K-2 data are used solely for outlier identification with K-2. Lastly, CTE and AP assessments do not undergo outlier identification due to the various test taking patterns inherent with CTE and AP and the fact that these assessments have less uniformity in administration across the state than other statewide assessments. 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 that no scores are used if they were in fact errors in the data, and the approach for flagging a student score as an outlier is fairly conservative.

Considerations included in outlier detection are:

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

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

from the weighted combination of all the reference scores. Using this t-value, EVAAS can flag individual scores as outliers.

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

For low-end outliers, the rules are:

- The percentile of the score must be below 50.
- The t-value must be below -3.5 for EOGs and EOCs when determining the difference between the score in question and the weighted combination of reference scores (otherwise known as the comparison score). In other words, the score in question must be at least 3.5 standard deviations below the comparison score. For other assessments, the t-value must be below -4.0.
- 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 need to be at least 10 to 40 percentiles above the individual percentile score.

For high-end outliers, the rules are:

- The percentile of the score must be above 50.
- The t-value must be above 4.5 for EOGs and EOCs when determining the difference between the score in question and the reference group of scores. In other words, the score in question must be at least 4.5 standard deviations above the comparison score. For other assessments, the t-value must be above 5.0.
- The percentile of the comparison score must be below a certain value. This value depends on the position of the individual score in question but will need to be at least 30 to 50 percentiles below the individual percentile score. There must be at least three reference scores used to make the comparison score.