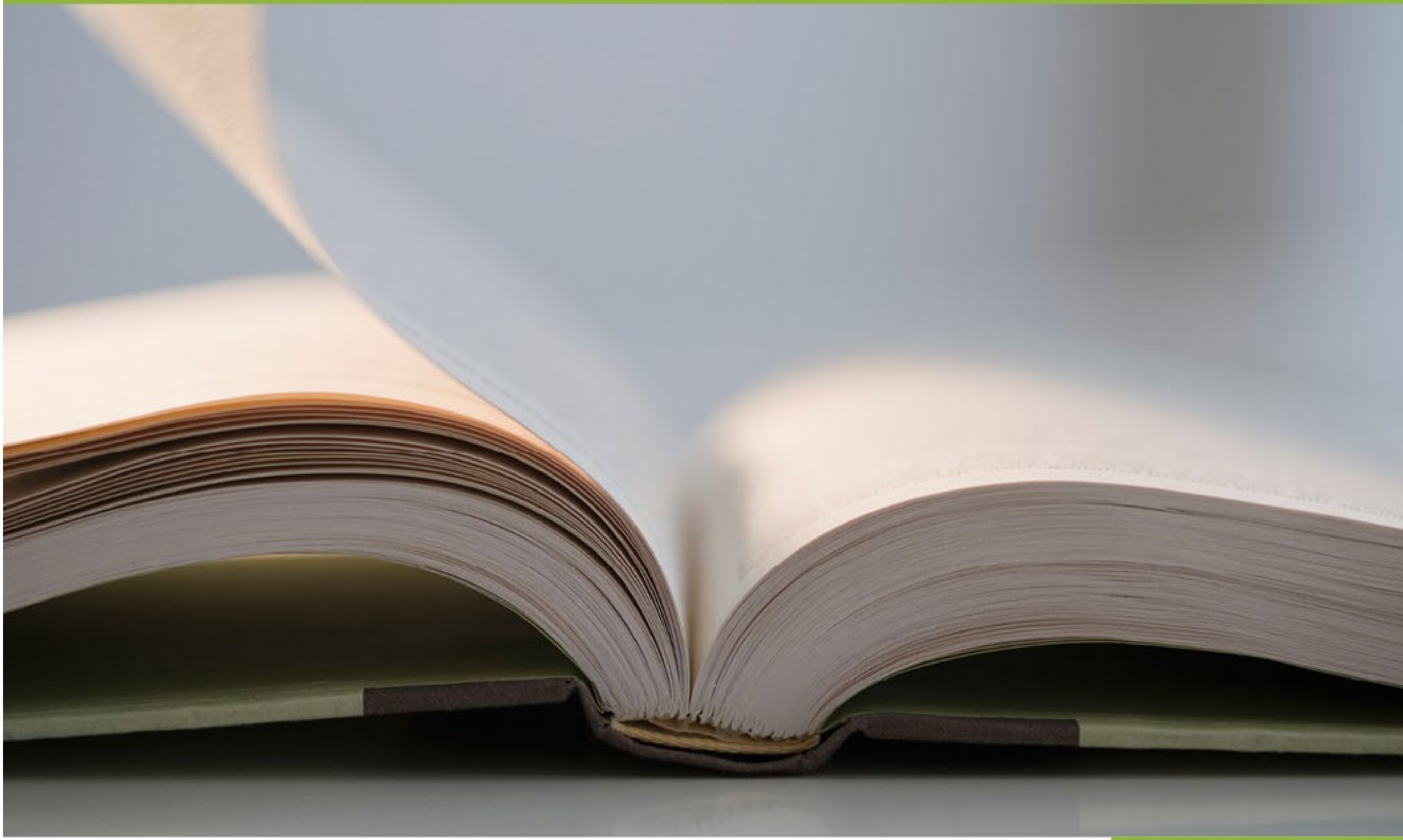
SAS® EVAAS

Statistical Models and Business Rules

Prepared for the Ohio Department of Education



Contents

1 Introduction to Ohio's Value-Added Reporting

The term "value-added" refers to a statistical analysis used to measure students' academic growth. Conceptually and as a simple explanation, value-added or growth measures are calculated by comparing the exiting achievement to the entering achievement for a group of students. Although the concept of growth is easy to understand, the implementation of a growth model is more complex.

First, there is not just one growth model; there are multiple growth models depending on the assessment, students included in the analysis, and level of reporting (district, school, or teacher). For each of these models, there are business rules to ensure the growth measures reflect the policies and practices selected by the State of Ohio.

Second, in order to provide reliable growth measures, growth models must overcome non-trivial complexities of working with student assessment data. For example, students do not have the same entering achievement; students do not have the same set of prior test scores; and all assessments have measurement error because they are estimates of student knowledge. EVAAS growth models have been in use and available to educators in states since the early 1990s. These growth models were among the first in the nation to use sophisticated statistical models that addressed these concerns.

Third, the growth measures are relative to students' expected growth, which is in turn determined by the growth that is observed within the actual population of Ohio test-takers in a subject, grade, and year. Interpreting the growth measures in terms of their distance from expected growth provides a more nuanced and statistically robust interpretation.

With these complexities in mind, the purpose of this document is to guide you through Ohio's value-added modeling based on the statistical models, business rules, policies, and practices selected by the State of Ohio and currently implemented by EVAAS. This document includes details and decisions in the following areas:

- Conceptual and technical explanations of analytic models
- Definition of expected growth
- Classifying growth into categories for interpretation
- Explanation of district, school, and teacher composites
- Input data
- Business rules

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

These reports are delivered through the EVAAS web application available at http://ohiova.sas.com. Although the underlying statistical models and business rules supporting these reports are sophisticated and comprehensive, the web reports are designed to be user-friendly and visual so that educators and administrators can quickly identify strengths and opportunities for improvement and then use these insights to inform curricular, instructional, and planning supports.

2 Statistical Models

2.1 Overview of Statistical Models

The conceptual explanation of value-added reporting is simple: compare students' exiting achievement with their entering achievement over two points in time. In practice, however, measuring student growth is more complex. Students start the school year at different levels of achievement. Some students move around and have missing test scores. Students might have "good" test days or "bad" test days. Tests, standards, and scales change over time. A simple comparison of test scores from one year to the next does not incorporate these complexities. However, a more robust value-added model, such as the one used in Ohio, can account for these complexities and scenarios.

Ohio's value-added models offer the following advantages:

- **The models use multiple subjects and years of data.** This approach minimizes the influence of measurement error inherent in all academic assessments.
- **The models can accommodate students with missing test scores.** This approach means that more students are included in the model and represented in the growth measures. Furthermore, because certain students are more likely to have missing test scores than others, this approach provides less biased growth measures than growth models that cannot accommodate students with missing test scores.
- **The models can accommodate tests on different scales.** This approach gives flexibility to policymakers to change assessments as needed without a disruption in reporting. It permits more tests to receive growth measures, particularly those that are not tested every year.
- **The models can accommodate team teaching or other shared instructional practices.** This approach provides a more accurate and precise reflection of student learning among classrooms.

These advantages provide robust and reliable growth measures to districts, schools, and teachers. This means that the models provide valid estimates of growth given the common challenges of testing data. The models also provide measures of precision along with the individual growth estimates taking into account all of this information.

Furthermore, because this robust modeling approach uses multiple years of test scores for each student and includes students who are missing test scores, EVAAS value-added measures typically have very low correlations with student characteristics. It is not necessary to make *direct* adjustments for student socioeconomic status or demographic flags because each student serves as their own control. In other words, to the extent that background 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:

> If 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.

Through this approach, the Ohio Department of Education (ODE) does not provide growth models to educators based on differential expectations for groups of students based on their backgrounds.

Based on Ohio's state assessment program, there are two approaches to providing district, school, and teacher growth measures.

- **The gain model (also known as the multivariate response model or MRM)** is used for tests given in consecutive grades, such as OST Mathematics and ELA assessments in grades 3–8.
- **The predictive model (also known as univariate response model or URM)** is used when a test is given in non-consecutive grades, such as OST Science assessments in grades 5 and 8 or any end-of-course tests.

There is another model, which is similar to the predictive model except that it is intended as an instructional tool for educators serving students who have not yet taken an assessment.

- **The projection model** is used for all assessments and provides a probability of obtaining a particular score or higher on a given assessment for individual students.

The following sections provide technical explanations of the models. The online Help within the EVAAS web application is available at https://ohiova.sas.com, and it provides educator-focused descriptions of the models.

2.2 Gain Model

2.2.1 Overview

The gain model measures growth between two points in time for a group of students; this is the case for tests given in consecutive grades such as OST Mathematics and ELA assessments in grades 3–8. **More specifically, the gain model measures the change in relative achievement for a group of students based on the statewide achievement from one year to the next.** Expected growth means that students maintained their relative achievement among the population of test-takers, and more details are available in Section 3.

There are three separate analyses for EVAAS reporting based on the gain model: 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 characteristic of districts and most schools. The teacher model uses a version adapted to the smaller numbers of students typically found in teachers' classrooms.

In statistical terms, the gain model is known as a linear mixed model and can be further described as a multivariate repeated measures model. These models have been used for value-added analysis for almost three decades, but their use in other industries goes back much further. These models were developed to use in fields with very large longitudinal data sets that tend to have missing data.

Value-added experts consider the gain model to be among one of the most statistically robust and reliable models. The references below include foundational studies by experts from RAND Corporation, a non-profit research organization:

- On the **choice of a complex value-added model**: McCaffrey, Daniel F., and J.R. Lockwood. 2008. "Value-Added Models: Analytic Issues." Prepared for the National Research Council and the

National Academy of Education, Board on Testing and Accountability Workshop on Value-Added Modeling, Nov. 13-14, 2008, Washington, DC.

- On the **advantages of the longitudinal, mixed model approach**: Lockwood, J.R. and Daniel McCaffrey. 2007. “Controlling for Individual Heterogeneity in Longitudinal Models, with Applications to Student Achievement.” *Electronic Journal of Statistics* 1:223-252.
- On the **insufficiency of simple value-added models**: McCaffrey, Daniel F., B. Han, and J.R. Lockwood. 2008. “From Data to Bonuses: A Case Study of the Issues Related to Awarding Teachers Pay on the Basis of the Students' Progress.” Presented at Performance Incentives: Their Growing Impact on American K-12 Education, Feb. 28-29, 2008, National Center on Performance Incentives at Vanderbilt University.

2.2.2 Why the Gain Model is Needed

A common question is why growth cannot be measured with a simple gain model that measures the difference between the current year’s scores and prior year’s scores for a group of students. The example in Figure 1 illustrates why a simple approach is problematic.

Assume that 10 students are given a test in two different years with the results shown in Figure 1. The goal is to measure academic growth (gain) from one year to the next. Two simple approaches are to calculate the mean of the differences *or* to calculate the differences of the means. When there is no missing data, these two simple methods provide the same answer (5.8 on the left in Figure 1). When there is missing data, each method provides a different result (6.9 versus 4.6 on the right in Figure 1).

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	**50.0**	**55.8**	**5.8**
Difference between Current and Previous Score Means			**5.8**

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

A more sophisticated model can account for the missing data and provide a more reliable estimate of the gain. As a brief overview, the gain model uses the correlation between current and previous scores in the non-missing data to estimate means for all previous and current scores as if there were no missing data. It does this without explicitly imputing values for the missing scores. The difference between these two estimated means is an estimate of the average gain for this group of students. In this example, the gain model calculates the estimated difference to be 5.8. Even in a small example such as this, the estimated difference is much closer to the difference with no missing data than either measure obtained by the mean of the differences (6.9) or the difference of the means (4.6). This method of estimation has been shown, on average, to outperform both of the simple methods.[1] This small example only considered two grades and one subject for 10 students. Larger data sets, such as those used in the actual value-added analyses for the state, provide better correlation estimates by having more student data, subjects, and grades. In turn, these provide better estimates of means and gains.

This simple example illustrates the need for a model that will accommodate incomplete data sets, which all student testing sets are. The next few sections provide more technical details about how the gain model calculates student growth.

2.2.3 Common Scale in the Gain Model

2.2.3.1 Why the Model Uses Normal Curve Equivalents

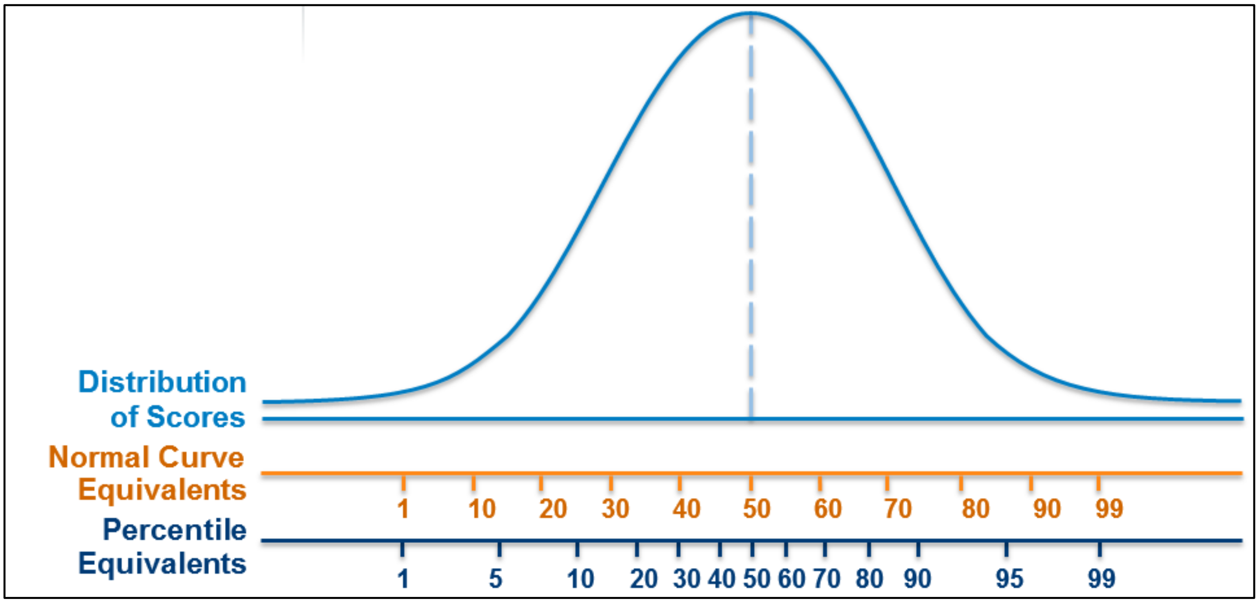
The gain model estimates academic growth as a "gain," or the difference between two measures of achievement from one point in time to the next. For such a difference to be meaningful, the two measures of achievement (that is, the two tests whose means are being estimated) must measure academic achievement on a common scale. Even for some vertically scaled tests, there can be different growth expectations for students based on their entering achievement. A reliable alternative to whether tests are vertically scaled is to convert scale scores to normal curve equivalents (NCEs).

An NCE distribution is similar to a percentile one. Both distributions provide context as to whether a score is relatively high or low compared to the other scores in the distribution. In fact, NCEs are constructed to be equivalent to percentile ranks at 1, 50, and 99 and to have a mean of 50 and standard deviation of approximately 21.063.

However, NCEs have a critical advantage over percentiles for measuring growth: NCEs 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. This difference between the distributions is evident below in Figure 2.

[1] See, for example, S. Paul Wright, "Advantages of a Multivariate Longitudinal Approach to Educational Value-Added Assessment without Imputation," Paper presented at National Evaluation Institute, 2004. Available online at https://evaas.sas.com/support/EVAAS-AdvantagesOfAMultivariateLongitudinalApproach.pdf.

Figure 2: Distribution of Achievement: Scores, NCEs, and Percentile Rankings



Furthermore, although percentile ranks are usually truncated below 1 and above 99, NCEs can range below 0 and above 100 to preserve the equal-interval property of the distribution and to avoid truncating the test scale. In a typical year among Ohio's state assessments, the average maximum NCE is approximately 125. Although the gain model does not use truncated values, which could create an artificial floor or ceiling in students' test scores, the web reporting shows NCEs as integers from 1 to 99 for display purposes.

2.2.3.2 Sample Scenario: How to Calculate NCEs in the Gain Model

The NCE distributions used in the gain model are based on a reference distribution of test scores in Ohio. This reference distribution is the distribution of scores on a state-mandated test for all students in a given year. By definition, the mean (or average) NCE score for the reference distribution is 50 for each grade and subject. For identifying the other NCEs, the gain model uses a method that does not assume that the underlying scale is normal. This method ensures an equal-interval scale, even if the testing scales are not normally distributed.

Table 1 provides an example of how the gain model converts scale scores to NCEs. The first five columns of the table are based on a tabulated distribution of about 130,000 test scores from Ohio data. In a given subject, grade, and year, the tabulation shows, for each given score, the number of students who scored that score ("Frequency") as well as the percentage ("Percent") that frequency represents out of the entire population of test-takers. The table also tabulates the "Cumulative Frequency as the number of students who made that score or lower and its associated percentage ("Cumulative Percent").

The next column, "Percentile Rank," converts each score to a percentile rank. As a sample calculation using the data in Table 1 below, the score of 425 has a percentile rank of 45.2. The data show that 43.5% of students scored *below* 425 while 46.9% of students scored *at or below* 425. To calculate percentile ranks with discrete data, the usual convention is to consider half of the 3.4% reported in the Percent column to be "below" the cumulative percent and "half" above the cumulative percent. To calculate the percentile rank, half of 3.4% (1.7%) is added to 43.5% from Cumulative Percent to give you a percentile rank of 45.2, as shown in the table.

Table 1: Converting Tabulated Test Scores to NCE Values

Score	Frequency	Cumulative Frequency	Percent	Cumulative Percent	Percentile Rank	Z-Score	NCE
418	3,996	48,246	3.1	36.9	35.4	-0.375	42.10
420	4,265	52,511	3.3	40.2	38.5	-0.291	43.87
423	4,360	56,871	3.3	43.5	41.8	-0.206	45.66
425	4,404	61,275	3.4	46.9	45.2	-0.121	47.46
428	4,543	65,818	3.5	50.4	48.6	-0.035	49.27
430	4,619	70,437	3.5	53.9	52.1	0.053	51.12
432	4,645	75,082	3.6	57.4	55.7	0.143	53.00

NCEs are obtained from the percentile ranks using the normal distribution. The table of the standard normal distribution (found in many textbooks[2]) or computer software (for example, a spreadsheet) provides 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. As mentioned above, the NCE distribution is scaled so that NCEs exactly match the percentile ranks at 1, 50, and 99. To do this, each Z-score is multiplied by approximately 21.063 (the standard deviation on the NCE scale) and then 50 (the mean on the NCE scale) is added.

With the test scores converted to NCEs, growth is calculated as the difference from one year and grade to the next in the same subject for a group of students. This process is explained in more technical detail in the next section.

2.2.4 Technical Description of the Gain Model

2.2.4.1 Definition of the Linear Mixed Model

As a linear mixed model, the gain model for district, school, and teacher value-added reporting is represented by the following equation in matrix notation:

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

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

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

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

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

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

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

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

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

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

where R is the $m \times m$ matrix that reflects the amount of variation in and 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 amount of variation in and the correlation among the random effects. If (v, ϵ) are normally distributed, the joint density of (y, v) is maximized when β has value b and v has value u given by the solution to the following equations, known as Henderson's mixed model equations:[3]

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

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

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

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

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

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

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

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

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

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

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

[3] McLean, Robert A., William L. Sanders, and Walter W. Stroup (1991). "A Unified Approach to Mixed Linear Models." The American Statistician, Vol. 45, No. 1, pp. 54-64.

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

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

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

4. With G and R known, the solution for the fixed effects is equivalent to generalized least squares, and if v and ϵ are multivariate normal, then the solutions for β and v are maximum likelihood.
5. If G and R are not known, then as the estimated G and R approach the true G and R, the solution approaches the maximum likelihood solution.
6. If v and ϵ are not multivariate normal, then the solution to the mixed model equations still provides the maximum correlation between v and u.

2.2.4.2 District and School Models

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

Algebraically, the district gain model 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_{jkld} 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 gain model is represented as:

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

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

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

Solving the mixed model equations for the district or school gain model produces a vector b that contains the estimated mean score for each school (in the school model), subject, grade, and year. To obtain a value-added measure of average student growth, a series of computations can be done using the students from a school in a particular year and their prior and current testing data. The model produces means in each subject, grade, and year that can be used to calculate differences in order to obtain gains. Because students 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 students who existed in the current year of that school. Therefore, mobility is taken into account within the model. Growth of students is computed using all students in each school including those that 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) as described above.

2.2.4.3 Teacher Model

The teacher estimates use a more conservative statistical process to lessen the likelihood of misclassifying teachers. Each teacher's growth measure is assumed to be equal to the state average in a specific year, subject, and grade until the weight of evidence pulls them either above or below that state average. The model also accounts for the percentage of instructional responsibility the teacher has for each student during the course of each school year. Furthermore, the teacher model is "layered," which means that:

- Students' performance with both their current and previous teacher effects are incorporated.
- For each school year, the teacher estimates are based students' testing data collected over multiple previous years.

Each element of the statistical model for teacher value-added modeling provides an additional level of protection against misclassifying each teacher estimate.

To allow for the possibility of many teachers with relatively few students per teacher, the gain model enters teachers as random effects via the Z matrix in the linear mixed model. The X matrix contains a column for each subject, grade, and year, and the b vector contains an estimated state mean score for each subject, grade, and year. The Z matrix contains a column for each subject, grade, year, and teacher, and the u vector contains an estimated teacher effect for each subject, grade, year, and 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, and 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 the j^{th} subject in the k^{th} grade in the l^{th} year, and I is an identity matrix.

Algebraically, the teacher model is represented as:

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

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. $\tau_{jk^*l^*t}$ is the teacher effect of the t^{th} teacher in the j^{th} subject in grade k^* in year l^*. The complexity of the parenthesized term containing the teacher effects is due to two factors. First, in any given subject, grade, and 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, and year, denoted by $T_{ijk^*l^*}$. $\tau_{jk^*l^*t}$ is the effect of the t^{th} teacher. $w_{ijk^*l^*t}$ is the fraction of the i^{th} student's instructional time claimed by the t^{th} teacher. Second, as mentioned above, this model allows teacher effects to accumulate over time. The outer (leftmost) summation accumulates teacher effects not only for the current (subscripts k and l) but also over previous grades and years (subscripts k^* and l^*) in the same subject. Because of this accumulation of teacher effects, this type of model is often called the "layered" model.

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

Table 2 illustrates how the Z matrix is encoded for three students who have three different scenarios of teachers during grades 3, 4, and 5 in two subjects, Math (M) and Reading (R). Teachers are identified by the letters A–F.

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

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

Eric's scenario illustrates team teaching. In grade 3 Reading, Eric received an equal amount of instruction from teachers A and B. The entries in the Z matrix indicate each teacher's contribution, 0.5

for each teacher. In grade 5 Math, however, Eric was taught by both teachers E and F, but they did not make an equal contribution. Teacher E claimed 80% responsibility, and teacher F claimed 20%.

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

Table 2: Encoding the Z Matrix

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

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).

2.2.4.4 Student Groups Model

For accountability purposes, the gain model provides district and school growth measures for their accountable students included in a specific student group. More details about accountable and tested students are available in Section 7.3.3.

In this student group analysis, expected growth is the same as in the overall students' analysis. In other words, expected growth is based on all students since the NCE mapping is based on all students, not just those in a specific student group. Furthermore, the estimated covariance parameters are used from the overall students' analysis when calculating the value-added measures.

For state accountability, the student groups include the following district and school measures:

- Overall (or all students)
- Gifted
- Students with Disabilities

For federal/ESSA accountability, the student groups include the following district and school measures:

- American Indian or Alaskan Native
- Asian or Pacific Islander
- Black, Non-Hispanic
- Community School Closure
- Economically Disadvantaged
- Gifted Students
- Hispanic
- Limited English Proficiency
- Multiracial
- Students with Disabilities
- White, Non-Hispanic

For additional informational purposes, reporting for student groups also includes the following:

- Student groups at community schools in accordance with state policies.

The business rules for identifying students in each group are provided in Section 7.3.4.

2.2.4.5 Drop-Out Recovery Program Model

Growth measures are required for dropout recovery programs, and given the unique nature of student enrollment, student grade, and student testing in these programs, ODE has customized the value-added modeling and data inputs for a more meaningful growth measure. The purpose of this section is to give a technical overview of this customized approach for the schools participating in these programs.

At the dropout recovery programs, students take assessments upon entering the program and again after they have received at least 84 days of instruction.

The tests that are used in this analysis were selected by ODE through a competitively bid contract. One property of the selected assessments is that they are computer adaptive since the grade level can be difficult to determine for some students. More information about these assessments can be found on ODE’s website.

The value-added model for dropout recovery programs is similar to the multivariate response model (MRM) or gain model currently used for OST Mathematics and Reading in non-dropout recovery schools in the state. In less technical terms, growth is measured through a gain-based approach using the two test scores in the same subject within a given year. The growth measure itself is the estimated *change in achievement* for a group of students with a specific program relative to the norm referenced population for that subject and grade. This measure considers the entering achievement of the group of students.

As a first step, the distribution of scores for a subject/grade/test window are mapped to a normal curve equivalent distribution using the norm data provided by the test vendor. This norm information is from a typical 10th grader testing in April. This does not assume anything about the achievement of individuals included in the analysis; it only puts them on a referenced curve of achievement to be able to compare their scores over time with an equal expectation of growth. The average score for the first test of a specific program is compared to its average score for the second test. The expected growth is that students will maintain their achievement levels between the two tests relative to the norm-referenced population, and the growth measure is the difference between the two achievement levels.

To determine whether the growth measure represents significantly more or less progress than the expected growth, a growth index is then calculated by dividing the growth measure by its standard error. The growth index is categorized into three levels: Exceeds Standards, Meets Standards, and Does Not Meet Standards. Multi-year growth measures are also reported where sufficient data exist. Prior to the 2019 reporting, the norms used were different. They represented a complete school year instead of the 84 days of instruction that is currently used. The difference in the interpretation from the OST growth measure is that the non-dropout recovery schools are measuring whether students maintained their same relative position in the distribution of statewide student achievement from one year to the next. The dropout recovery schools are using a national norm assessment and measuring whether students maintained their same relative position in that national norm referenced group from the initial test at the time of program entry to the second assessment at least 13 weeks later.

2.2.5 Modeling Adjustments to 2020-21 Growth Measures to Accommodate Missing 2019-20 Data

2.2.5.1 Overview

In spring 2020, the COVID-19 pandemic required schools to close early and cancel statewide summative assessments. As a result, scores are not available for OST assessments based on the 2019-20 school year, and the 2020-21 EVAAS reporting does not include 2019-20 test scores. In essence, the 2020-21 EVAAS reporting based on the gain model represents a two-year growth measure, measuring the change in achievement from the 2018-19 school year to the 2020-21 school year.

To conceptualize what the 2020-21 growth measures mean for districts and schools, Table 3 provides the average achievement level for the students testing at a sample school. As a cohort of students moves from one grade to the next, their achievement level can be tracked along a diagonal line. For example, Table 3 shows that the achievement level of Grade 5 students in Year 2 is 25 NCEs and then changes to 36 NCEs when this cohort of students is in Grade 6 in Year 3.

Table 3: Average Achievement in NCEs by Grade and Year for Sample School

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

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

In reality (not the computationally ideal situation described above), the gain model calculates means by accounting for missing student scores.

The achievement level reported for Grade 6 in Year 3 is an average based on the students' prior test scores from other schools. This is relevant for the lowest grade in a school, often Grade 6, because there is no mean *at that school* for the previous grade and year.

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

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

To determine the feasibility of this approach, the cumulative gain could be compared to the sum of the one-year gains based on a model with Year 2 data. This would be (36 NCEs – 25 NCEs) + (25 NCEs – 14 NCEs), which would be 11 NCEs + 11 NCEs, or 22 NCEs. The ideal case is that the cumulative two-year gain and the sum of the one-year gains are the same. In practice, they might differ due to lack of information about missing student data.

At ODE's request, SAS conducted research to simulate the impact of a missing year of data on results. This research compared growth results from different models to assess the impact of a missing year of data. The high-level findings are below:

- A comparison between the standard, one-year gain model used for consecutive grade-given tests and gain model results with a missing prior year of data illustrates positive correlations for all individual grades and subjects for districts (0.48 to 0.71), schools (0.49 to 0.71), and teachers (0.68 to 0.81). In this case, one of the comparisons is essentially based on a two-year gain, whereas the other is based on a one-year gain.
- To provide a more comparable analysis, the analysis also compared a sum of single-year 2017-18 and 2018-19 growth measures based on the gain model *with* 2017-18 test scores to the 2018-19 growth measures based on the gain model *without* 2017-18 test scores. The results from these two models are very similar, with a correlation above 0.99 across all district and

school results for ELA and Math for all individual subject/grades. This finding is not surprising as both models represent a growth measure from a two-year period. Because teachers tend to teach different students from year-to-year, this comparison cannot be made at the teacher level. It also did not include the first grade of a school since there were not two consecutive gains to sum together for a comparison.

Although a precise definition varies, a typical interpretation of a positive correlation is that a weak relationship is between 0.10 and 0.30; a moderate relationship is between 0.30 and 0.50; and a strong relationship is above 0.50. This suggests that gain model results, while strongly correlated with one-year gains with a prior year of data missing, can be most accurately interpreted as a two-year gain. To the extent that districts, schools, and teachers served students in the current year but not the prior year, the interpretation of the growth measure can include students' experiences with a different district, school, or teacher in the previous year.

It is important to note that these simulations were focused solely on the missing year of data and the appropriateness of estimating two-year gains; they did not estimate the pandemic's impact on student learning in districts, schools, and classrooms.

2.3 Predictive Model

2.3.1 Overview

Tests that are not given in consecutive grades require a different modeling approach from the gain model. The predictive model is used for such assessments in Ohio. **The predictive model is a regression-based model where growth is a function of the difference between students' expected scores with their actual scores.** Expected growth is met when students with a district, school, or teacher made the same amount of growth as students with the average district, school, or teacher.

Like the gain model, there are three separate analyses for EVAAS reporting based on the predictive model: one each for districts, schools, and teachers. The district and school models are essentially the same, and the teacher model includes accommodations for team teaching and other shared instruction.

Regression models are used in virtually every field of study, and their intent is to identify relationships between two or more variables. When it comes to measuring growth, regression models identify the relationship between prior test performance and actual test performance for a given course. In more technical terms, the predictive model is known as the univariate response model (URM), a linear mixed model and, more specifically, an analysis of covariance (ANCOVA) model.

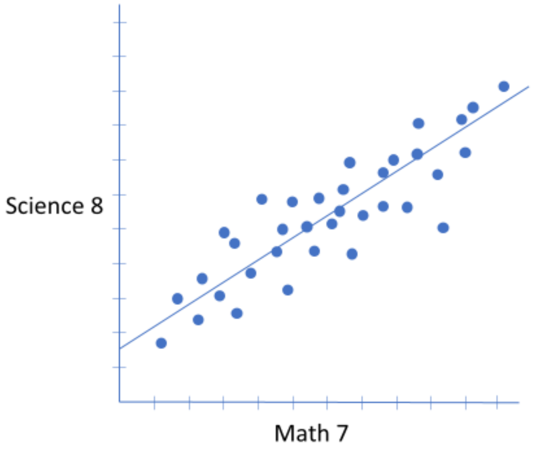
The key advantages of the predictive model can be summarized as follows:

- It minimizes the influence of measurement error and increases the precision of predictions by using multiple prior test scores as predictors for each student.
- It 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 as predictors of the response variable in any subject and grade.
- It allows educators to benefit from all tests, even when tests are on differing scales.
- It accommodates teaching scenarios where more than one teacher has responsibility for a student's learning in a specific subject, grade, and year.

2.3.2 Conceptual Explanation

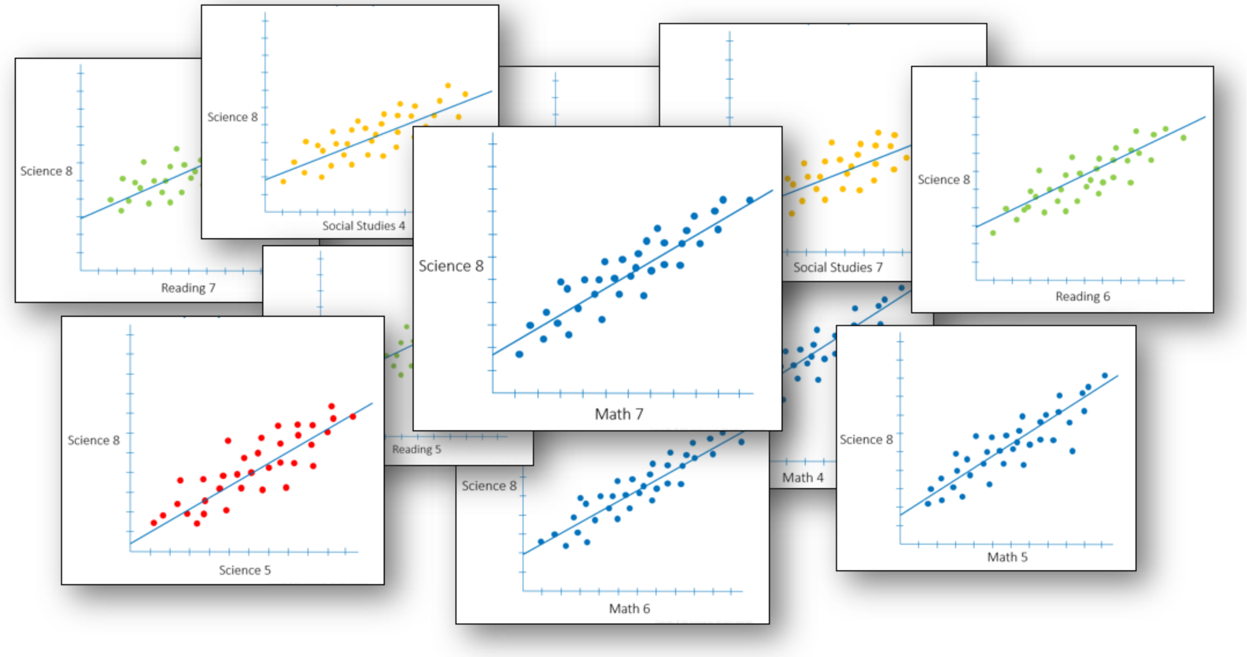
As mentioned above, the predictive model is ideal for assessments given in non-consecutive grades, such as OST Science tests in grades 5 and 8, or the high school end-of-course tests, such as Algebra I. Consider all students who tested in OST Science in grade 8 in a given year. The gain model is not possible since there isn't a Science test in the immediate prior grade. However, these students might have a number of prior test scores in OST Math and ELA in grades 3–7 and OST Science in grade 5. These prior test scores have a relationship with OST Science, meaning that how students performed on these tests can predict how the students perform on OST Science in grade 8. The growth model does not assume what the predictive relationship will be; instead, the actual relationships observed by the data define the relationships. This is shown in Figure 3 below where each dot represents a student's prior score on OST Math 7 plotted with their score on OST Science 8. The best-fit line indicates how students with a certain prior score on OST Math 7 tend to score, on average, on OST Science 8. This illustration is based on one prior test; the predictive model uses many prior test scores from different subjects and grades.

Figure 3: Test Scores from One Assessment Have a Predictive Relationship to Test Scores from Another Assessment



Some subjects and grades will have a greater relationship to OST Science in grade 8 than others; however, the other subjects and grades still have a predictive relationship. For example, prior Math scores might have a stronger predictive relationship to OST Science in grade 8 than prior ELA scores, but how a student reads and performs on the OST ELA test typically provides an idea of how we might expect a student to perform on average on OST Science. This is shown in Figure 4 below, where there are a number of different tests that have a predictive relationship with OST Science in grade 8. All of these relationships are considered together in the predictive model, with some tests weighted more heavily than others.

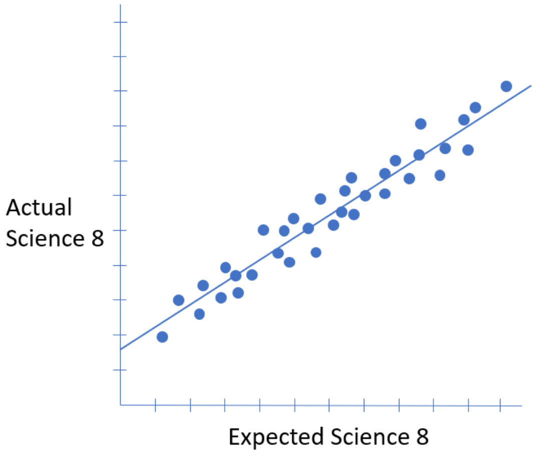
Figure 4: Relationships Observed in the Statewide Data Inform the Predictive Model



Note that the prior test scores do not need to be on the same scale as the assessment being measured for student growth. Just as height (reported in inches) and weight (reported in pounds) can predict a child's age (reported in years), the growth model can use test scores from different scales to find the predictive relationship.

Each student receives an expected score based on their own prior testing history. In practical terms, the expected score represents the student's entering achievement because it is based on all prior testing information to date. Figure 5 below shows the relationship between expected and actual scores for a group of students.

Figure 5: Relationship Expected Score and Actual Score for Selected Subject and Grade



The expected scores can be aggregated to a specific district, school, or teacher and then compared to the students' actual scores. In other words, the growth measure is a function of the difference between the exiting achievement (or average actual score) and the entering achievement (or average expected score) for a group of students. Unlike the gain model, the actual score and expected score are reported in the scaling units of the test rather than NCEs.

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

The predictive model has similar approaches for districts and schools and a slightly different approach for teachers that accounts for shared instructional responsibility. The approach is described briefly below, with more details following.

- The score to be predicted serves as the response variable (y, the dependent variable).
- The covariates (x terms, predictor variables, explanatory variables, independent variables) are scores on tests the student has taken in previous years from the response variable.
- There is a categorical variable (class variable, grouping variable) to identify the district, school, or teacher(s) from whom the student received instruction in the subject, grade, and year of the response variable (y).

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

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

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

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

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, because the predictive model is an ANCOVA model, the estimated parameters are pooled within group (district, school, or teacher). The strategy for dealing with missing predictors is to estimate the joint covariance matrix (call it C) of the response and the predictors. Let C be partitioned into response (y) and predictor (x) partitions, that is,

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

Note that *C* in equation (16) is not the same as *C* in equation (4). This matrix is estimated using the EM (expectation maximization) algorithm for estimating covariance matrices in the presence of missing data available in SAS/STAT® (although no imputation is actually used). It should also be noted that, due to this being an ANCOVA model, *C* is a pooled-within group (district, school, or teacher) covariance matrix. This is accomplished by providing scores to the EM algorithm that are centered around group means (i.e., the group means are subtracted from the scores) rather than around grand means. Obtaining C is an iterative process since group means are estimated within the EM algorithm to accommodate missing data. Once new group means are obtained, another set of scores is fed into the EM algorithm again until C converges. This overall iterative EM algorithm is what accommodates the two difficulties mentioned above. Only students who had a test score for the response variable in the most recent year and who had at least three predictor variables are included in the estimation. Given such a matrix, the vector of estimated regression coefficients for the projection equation (15) can be obtained as:

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

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

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

Once the parameter estimates for the prediction equation have been obtained, predictions can be made for any student with any set of predictor values as long as that student has a minimum of three prior test scores. This is to avoid bias due to measurement error in the predictors.

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

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

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

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

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

2.3.4 Modeling Adjustments to 2020-21 Growth Measures to Accommodate Missing 2019-20 Data

In spring 2020, the COVID-19 pandemic required the cancellation of statewide summative assessments. As a result, scores are not available for the OST assessments based on the 2019-20 school year, and the 2020–21 EVAAS reporting does not include 2019-20 test scores in the models.

The predictive model is used to measure growth for assessments given in non-consecutive grades, such as OST Science assessments in grades 5 and 8 or any end-of-course tests. Because these assessments are not administered every year, it is always possible that students do not have any test scores in the immediate prior year. The model can provide a robust estimate of students' entering achievement for the course by using all other available test scores from other subjects, grades, and years.

In other words, the predictive model did not require any technical adaptations to account for the missing year of data; any predictors from the previous year were just excluded from the model.

As with the gain model, SAS conducted simulation research comparing actual 2018-19 results from the predictive model to 2018-19 results that removed the 2017-18 prior scores as predictors. The correlations between these results ranged from 0.92 to 0.98 at the district level, 0.93 to 0.97 at the school level, and 0.93 to 0.98 at the teacher level. This illustrates that results from the predictive model with a missing prior year of data are strongly correlated with the actual results without a missing year of data.

As with the gain model, it is important to note that these simulations were focused solely on the missing year of data and did not estimate the pandemic's impact on student learning in districts, schools, and classrooms.

2.4 Projection Model

2.4.1 Overview

The longitudinal data sets used to calculate growth measures for groups of students can also provide individual student projections to future assessments. A projection is reported as a probability of obtaining a specific score or above on an assessment, such as a 70% probability of scoring Proficient or above on the next summative assessment. The probabilities are based on the students' own prior testing history as well as how the cohort of students who just took the assessment performed. Due to the pandemic, the projections to the assessments in the 2021-22 school year are based on cohort of students who took the assessment in the 2018-19 school year. Projections are available for state assessments as well as to college readiness assessments.

[4] For more information about shrinkage estimation, see, for example, Ramon C. Littell, George A. Milliken, Walter W. Stroup, Russell D. Wolfinger, and Oliver Schabenberger, *SAS for Mixed Models, Second Edition* (Cary, NC: SAS Institute Inc., 2006). Another example is Charles E. McCulloch, Shayle R. Searle, and John M. Neuhaus, *Generalized, Linear, and Mixed Models, Second Edition* (Hoboken, NJ: John Wiley & Sons, 2008).

Projections are useful as a planning resource for educators, and they can inform decisions around enrollment, enrichment, remediation, counseling, and intervention to increase students' likelihood of future success.

2.4.2 Technical Description

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 predictive model applied at the school level described in Section 2.3.3. In the projection model, the score to be projected serves as the response variable (y), the covariates (x terms) 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, and year of the response variable (y). Algebraically, the model can be represented as follows for the i^{th} student.

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

The μ terms are means for the response and the predictor variables. α_j is the school effect for the j^{th} school, the school attended by the i^{th} student. The β terms are regression coefficients. Projections to the future are made by using this equation with estimates for the unknown parameters (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

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

The reason for the "±" before the $\hat{\alpha}_j$term is that since the projection is to a future time, the school that the student will attend is unknown, so this term is usually omitted from the projections. This is equivalent to setting $\hat{\alpha}_j$ to zero, that is, to assuming that the student encounters the "average schooling experience" in the future.

Two difficulties must be addressed to implement the projections. First, not all students will have the same set of predictor variables due to missing test scores. Second, because this is an ANCOVA model with a school effect i, the regression coefficients must be "pooled-within-school" regression coefficients. The strategy for dealing with these difficulties is the same as described in Section 2.3.3 using equations (16), (17), and (18) 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 least Proficient on a future end-of-grade test or the probability of scoring at least an established college readiness benchmark. The probability is calculated as the area above the benchmark cutoff score using a normal distribution with its mean equal to the projected score and its standard deviation equal to the standard error of the projected score as described below. Φ represents the standard normal cumulative distribution function.

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

2.5 Outputs from the Models

This section outlines the model outputs that are typically provided. Some outputs based on the 2020-21 reporting are not provided due to the pandemic, and these changes are noted below.

2.5.1 Gain Model

The gain model is used for courses where students test in consecutive grade-given tests. As such, **the gain model uses OST in Mathematics and English Language Arts in grades 3–8 to provide district, school, and teacher growth measures in the following content areas:**

- OST Mathematics in grades 4–8 (grade 4 not available for 2020-21)
- OST English Language Arts in grades 4–8

In addition to the mean scores and mean gain for an individual subject, grade and year, the gain model can also provide composites across subjects and across grades, such as those displayed on accountability reports. Multi-year composites for up to three years are also possible, though these are not included for the 2021-22 reporting.

In general, these are all different forms of linear combinations of the fixed effects (and random effects for the teacher model), and their estimates and standard errors are computed in the same manner described above in equations (5) and (6) for district and school models and in equations (9) and (10) for the teacher model.

Collectively, the different models provide metrics for a variety of purposes within the State of Ohio. They are summarized in the list below, and more details about the difference between accountable and tested students are in Section 7.3.3:

- District growth measures based on accountable students
  - Overall students
  - Gifted students
  - Students with disabilities
  - ESSA student groups, including White, Black, Asian/Pacific Islander, American Indian, Multi-Racial, Hispanic, English Learners, and Economically Disadvantaged
- District growth measures based on tested students
  - Overall students
- School growth measures based on accountable students
  - Overall students
  - Gifted students
  - Students with disabilities
  - ESSA student groups, including White, Black, Asian/Pacific Islander, American Indian, Multi-Racial, Hispanic, English Learners, and Economically Disadvantaged
  - Community school closure students (not available for 2020-21)
- School growth measures based on tested students
  - Overall students
- Teacher growth measures based on linked students

- Program growth measures
  - Drop-out recovery programs
  - STEM provider programs

Note that more details about district, school, and teacher composites across subjects, grades, and years are available in Section 5.

2.5.2 Predictive Model

The predictive model is used for courses where students test in non-consecutive grade-given tests. As such, **the predictive model provides growth measures for districts, schools, and teachers in the following content areas:**

- OST Science in grade 5 (not available for 2020-21) and 8
- OST EOC Algebra I
- OST EOC Mathematics I and II
- OST EOC Geometry
- OST EOC English Language Arts II
- OST EOC American US Government
- OST EOC American US History
- OST EOC Biology

In addition to the mean scores and growth measures for an individual subject, grade, and year, the predictive model can also provide multi-year average growth measures (up to three years) for each subject and grade or course.

Collectively, the different models provide metrics for a variety of purposes within the State of Ohio. They are summarized below, and more details about the difference between accountable and tested students are in Section 7.3.3:

- District growth measures based on accountable students
  - Overall students
  - Gifted students
  - Students with disabilities
  - ESSA student groups, including White, Black, Asian/Pacific Islander, American Indian, Multi-Racial, Hispanic, English Learners, and Economically Disadvantaged
- District growth measures based on tested students
  - Overall students
- School growth measures based on accountable students
  - Overall students
  - Gifted students
  - Students with disabilities
  - ESSA student groups, including White, Black, Asian/Pacific Islander, American Indian, Multi-Racial, Hispanic, English Learners, and Economically Disadvantaged
  - Community school closure criteria (not available for 2020-21)
- School growth measures based on tested students
  - Overall students

- Teacher growth measures based on linked students
- Program growth measures
  - STEM provider students

Note that while they were not calculated for 2021 due to considerations associated with the impact of the pandemic, more details about district, school and teacher composites across subjects, grades, and years are available in Section 5.

2.5.3 Projection Model

Projections are provided to future OST assessments as well as college readiness and Advanced Placement assessments. Projections provided in 2023 will use projection models that utilize the 2023 cohort of test takers.

Where applicable, projections are available for the following OST assessments:

- OST Mathematics 4-8
- OST English Language Arts 4-8
- OST Science grades 5 and 8
- OST EOCs (Algebra I, Mathematics I and II, ELA II, Geometry, Biology, American US Government, American US History)

More specifically, OST projections are provided to a student's next tested grade-level OST assessments based on that student's most recent tested grade, such as projections to grade 6 for students who most recently tested in grade 5. EOC projections for Algebra I and Mathematics I are provided starting with students who most recently tested in grade 7. EOC projections for English Language Arts II, Geometry, Mathematics II, and American US History are provided starting with students who most recently tested in grade 8. EOC projections for American US Government are provided starting with students who most recently tested in grade 9.

OST projections are made to the performance levels of Basic, Proficient, Accomplished and Advanced, and the individual cut scores depend on each subject and grade. OST projections are also made to the Competency Score for Graduation performance level for Algebra I and English Language Arts II.

ACT or SAT projections are provided to students who last tested in grades 7-11. ACT/SAT projections will be provided to the following cut scores based on the performance of the 2022 cohort:

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

AP projections are provided for students last tested in grades 7-11 or until they have taken the AP assessment in a given subject. Projections are provided for the likelihood of scoring 2 or higher, 3 or higher, or 4 or higher on the following AP assessments:

- AP Biology
- AP Calculus AB
- AP Chemistry

- AP Computer Science
- AP English Language and Composition
- AP English Literature and Composition
- AP Environmental Science
- AP European History
- AP Human Geography
- AP Macroeconomics
- AP Microeconomics
- AP Physics
- AP Psychology
- AP Statistics
- AP U.S. Government and Politics
- AP U.S. History
- AP World History

3 Expected Growth

3.1 Overview

Conceptually, growth is simply the difference between students' entering and exiting achievement. As noted in Section 2, zero represents "expected growth." Positive growth measures are evidence that students made *more* than the expected growth, and negative growth measures are evidence that students made *less* than the expected growth.

A more detailed explanation of expected growth and how it is calculated are useful for the interpretation and application of growth measures.

3.2 Technical Description

Both the gain and predictive models define expected growth based on the empirical student testing data; in other words, the model does not assume a particular amount of growth or assign expected growth in advance of the assessment being taken by students. Both models define expected growth within a year. This means that expected growth is always relative to how students' achievement has changed in the most recent year of testing rather than a fixed year in the past.

More specifically, **in the gain model, expected growth means that students maintained the same relative position with respect to the statewide student achievement that year. In the predictive model, expected growth means 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, and grade**.

For both models, the growth measures tend to be centered on expected growth every year with approximately half of the district/school/teacher estimates above zero and approximately half of the district/school/teacher estimates below zero.

A change in assessments or scales from one year to the next does not present challenges to calculating expected growth. Through the use of NCEs, the gain model converts any scale to a relative position, and the predictive model already uses prior test scores from different scales to calculate the expected score. When assessments change over time, expected growth is still based on the relative change in achievement from one point in time to another.

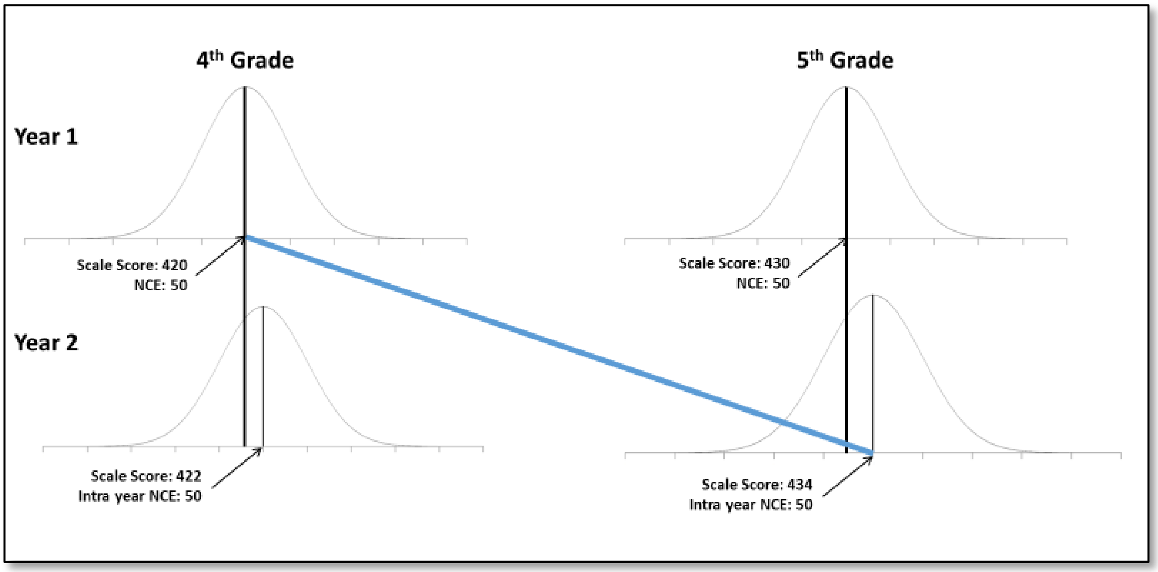
3.3 Illustrated Example

Figure 6 below provides a simplified example of how growth is calculated in the gain model when the state achievement increases. The figure has four graphs, each of which plot the NCE distribution of scale scores for a given year and grade. In 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 graph is shifted slightly to the right. The blue line shows what is required for students to make expected growth, which would be to maintain their position at the 50th NCE for grade 4 in Year 1 as they become grade 5 students in Year 2. The growth measure for these students is Year 2

NCE – Year 1 NCE, which would be 50 – 50 = 0. Similarly, if a group of students started at the 35th NCE, the expectation is that they would maintain that 35th NCE.

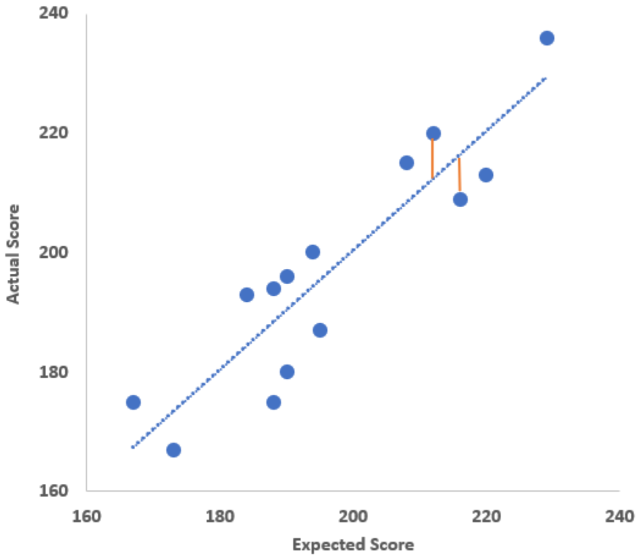
Note that the actual gain calculations are much more robust than what is presented here; as described in the previous section, the models can address students with missing data, team teaching, and all available testing history.

Figure 6: Intra-Year Approach Example for the Gain Model



In contrast, in the predictive model, expected growth uses actual results from the most recent year of assessment data and considers the relationships from the most recent year with prior assessment results. Figure 7 below provides a simplified example of how growth is calculated in the predictive model. The graph plots each student’s actual score with their expected score. Each dot represents a student, and a best-fit line will minimize the difference between all students’ actual and expected scores. Collectively, the best-fit line indicates what expected growth is for each student – given the student’s expected score, expected growth is met if the student scores the corresponding point on the best-fit line. Conceptually, with the best-fit line minimizing the difference between all students’ actual and expected scores, the growth expectation is defined by the average experience. Note that the actual calculations differ slightly since this is an ANCOVA model where the students are expected to see the average growth as seen by the experience with the average group (district, school, or teacher).

Figure 7: Intra-Year Approach Example for the Predictive Model



4 Classifying Growth into Categories

4.1 Overview

It can be helpful to classify growth into different levels for interpretation and context, particularly when the levels have statistical meaning. Ohio's growth model has three categories, which are defined by a range of values related to the growth measure and its standard error. These categories are known as growth indicators in the web application.

4.2 Use Standard Errors Derived from the Models

As described in the modeling approaches section, the growth model provides an estimate of growth for a district, school, or teacher in a particular subject, grade, and 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. It also takes into account shared instruction and team teaching. Standard error is a common statistical metric reported in many analyses and research studies because it yields important information for interpreting an estimate, in this case the growth measure relative to expected growth. 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 growth measure 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 they are truly effective) for high-stakes usage of value-added reporting.

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, and year could vary significantly among teachers, depending on the available data that is associated with their students, and it is another important protection for districts, schools, and teachers to incorporate standard errors to the value-added reporting.

4.3 Define Growth Indicators in Terms of Standard Errors

Common statistical usage of standard errors indicates the precision of an estimate and whether that estimate is statistically significantly different from an expected value. The growth reports use the standard error of each growth measure to determine the statistical evidence that the growth measure is different from expected growth.

4.3.1 Illustrated Examples of Categories

There are two ways to visualize how the growth measure and standard error relate to expected growth and how these can be used to create categories.

The first way is to frame the growth measure relative to its standard error and expected growth at the same time.

- **Light Blue** indicates that the growth measure is two standard errors or more above expected growth (0). This level of certainty is significant evidence of exceeding the standard for academic growth.
- **Green** indicates that the growth measure is less than two standard errors above expected growth (0) or up to two standard errors below expected growth (0). This is evidence of meeting the standard for academic growth.
- **Yellow** indicates that the growth measure is more than two standard errors below expected growth (0). This is significant evidence of not meeting the standard for academic growth.

The second way to frame the categories is to create a growth index, which is calculated as shown below:

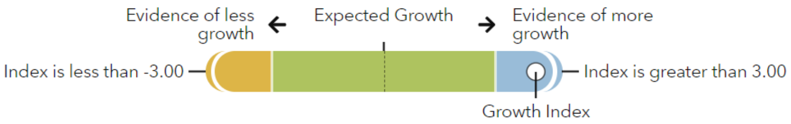
$$Growth\ Index = \frac{Growth\ Measure - Expected\ Growth}{Standard\ Error\ of\ the\ Growth\ Measure} \tag{23}$$

The growth index is similar in concept to a Z-score or t-value, and it communicates as a single metric the certainty or evidence that the growth measure is decidedly above or below expected growth. The growth index is useful when comparing value-added measures from different assessments or in different units, such as NCEs or scale scores. The categories can be established as ranges based on the growth index, such as the following:

- **Light Blue** indicates significant evidence that students made more growth than expected. The growth index is 2 or greater.
- **Green** indicates evidence that students made growth as expected. The growth index is between -2 and 2.
- **Yellow** indicates significant evidence that students made less growth than expected. The growth index less than -2.

This is represented in the effectiveness level bar in Figure 8, which is similar to what is provided in the District and School Value-Added reports in the EVAAS web application. The black dotted line represents expected growth. The color-coding within the bar indicates the range of values for the growth index within each category.

Figure 8: Sample Effectiveness Level Bar



It is important to note that these two illustrations provide users with the same information; they are simply presenting the growth measure, its standard error, and expected growth in different ways.

4.4 Define Growth Categories in Terms of Student-Level Standard Deviation of Growth

The student-level standard deviation of growth can be used to provide context about the magnitude of growth being made by a group of students. The standard deviation of the student-level distribution of growth is available for each year, subject, and grade. These standard deviations are specific to each year and the values for 2022-23 assessments are included in Table 4.

Table 4: Standard Deviation of Student-Level Distribution of Growth (2022-23)

Assessment	Grade	Year	Standard Deviation of Student-Level Distribution of Growth
OST Mathematics	4	2022-23	11.7889
OST Mathematics	5	2022-23	11.2548
OST Mathematics	6	2022-23	10.9568
OST Mathematics	7	2022-23	10.5842
OST Mathematics	8	2022-23	11.1228
OST English Language Arts	4	2022-23	13.6341
OST English Language Arts	5	2022-23	12.9166
OST English Language Arts	6	2022-23	12.3602
OST English Language Arts	7	2022-23	11.4619
OST English Language Arts	8	2022-23	11.2242
OST Science	5	2022-23	26.6365
OST Science	8	2022-23	25.4255
OST Algebra I	N/A	2022-23	17.6054
OST Biology	N/A	2022-23	15.2867
OST English Language Arts II	N/A	2022-23	15.3566
OST Geometry	N/A	2022-23	20.3132
OST American US Government	N/A	2022-23	10.9181
OST American US History	N/A	2022-23	15.9518
OST Mathematics I	N/A	2022-23	15.9223
OST Mathematics II	N/A	2022-23	21.2814

Dividing the growth measures by the standard deviation provides a value known as an "effect size," and it indicates the practical significance regarding the group of students and whether they met, exceeded, or fell short of expected growth. This effect size is used in the two-step categorization of accountability composites outlined in Section 5.

4.5 Rounding and Truncating Rules

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

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

5 Composite Growth Measures

A composite combines growth measures from different subjects, grades, and/or years. The sections below describe the calculation of composites for teachers, then schools and districts, and lastly principals.

5.1 Teacher Composites

5.1.1 Overview

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

- Beginning with the 2021-22 reporting, only single-year composites are calculated for teachers.
- The composite for teachers weights each subject and grade based on the FYE number of students used in that measure.

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

1. For growth measures based on the gain model, calculate composite index across subjects.
2. For growth measures based on the predictive model, calculate composite index across subjects.
3. Using both the gain and predictive model composite indices, calculate the composite index.

If a teacher does not have value-added measures from both the gain and predictive models, then the composite index would be based on the model for which the teacher does have reporting.

The following sections illustrate this process using value-added measures from a sample teacher, which are provided in Table 5.

Table 5: Sample Teacher Value-Added Information

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

5.1.2 Technical Description of the Composite Index based on Gain Model Measures

The composite index for the gain model growth measures is calculated by dividing the composite gain by its composite standard error. The calculations for each of these metrics are provided below.

5.1.2.1 Composite Gain Across Subjects

Growth measures from the gain model are in the same scale (NCEs), so the composite gain across subjects is a simple average gain where each growth measure is weighted according to the proportion of students linked to that gain. For our sample teacher, the total number of Full-Year Equivalent (FYE) students affiliated with growth measures from the gain model is 65 + 70, or 135 students. The ELA grade 8 growth measure would be weighted at 65/135, and the Math grade 8 growth measure would be weighted at 70/135.

The calculation of the composite gain across subjects based on the gain model is as follows:

$$Composite\ Gain = \frac{65}{135} ELA_8 + \frac{70}{135} Math_8 = \left(\frac{65}{135}\right)(-0.30) + \left(\frac{70}{135}\right)(3.80) = 1.83 \tag{24}$$

5.1.2.2 Composite Standard Error Across Subjects

As discussed in Section 4, the standard error is a measure of the statistical certainty in the growth measure that indicates whether an estimate is decidedly above or below expected growth. Standard errors can, and should, also be provided for the composite gains that have been calculated from a teacher's value-added gain estimate.

As background, statistical formulas are often more conveniently expressed as variances (see equation 6, for example), and this is the square of the standard error. Standard errors of composites can be calculated using variations of the general formula shown below. To maintain the generality of the formula, the individual estimates in the formula (think of them as value-added gains) are simply called X, Y, and Z. If there were more than or fewer than three estimates, the formula would change accordingly. As OST composites use proportional weighting according to the number of students linked to each value-added gain, each estimate is multiplied by a different weight: a, b, or c.

$$\begin{gathered} Var(aX + bY + cZ) = a^2 Var(X) + b^2 Var(Y) + c^2 Var(Z) \\ +2ab\ Cov(X,Y) + 2ac\ Cov(X,Z) + 2bc\ Cov(Y,Z) \end{gathered} \tag{25}$$

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

$$Cov(X,Y) = Correlation(X,Y)\sqrt{Var(X)}\sqrt{Var(Y)} \tag{26}$$

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

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

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

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 and grade 8 ELA.

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 and to compare it to the standard error calculated assuming (inappropriately) an extreme correlation of +1. Using the gain-based FYE weightings and standard errors and assuming total independence, the standard error would then be as follows:

$$\begin{aligned} \mathit{Composite\ Standard\ Error} &= \sqrt{\left(\frac{65}{135}\right)^2 (SE\ Read_8)^2 + \left(\frac{70}{135}\right)^2 (SE\ Math_8)^2} \\ &= \sqrt{\left(\frac{65}{135}\right)^2 (1.20)^2 + \left(\frac{70}{135}\right)^2 (1.50)^2} = 0.97 \end{aligned} \tag{27}$$

Assuming a correlation of +1, the standard error would then be as follows:

$$\begin{aligned} &\mathit{Composite\ Standard\ Error} \\ &= \sqrt{\left(\frac{65}{135}\right)^2 (SE\ Read_8)^2 + \left(\frac{70}{135}\right)^2 (SE\ Math_8)^2 + 2\left(\frac{65}{135}\right)\left(\frac{70}{135}\right)(SE\ Read_8)(SE\ Math_8)} \\ &= \sqrt{\left(\frac{65}{135}\right)^2 (1.20)^2 + \left(\frac{70}{135}\right)^2 (1.50)^2 + 2\left(\frac{65}{135}\right)\left(\frac{70}{135}\right)(1.20)(1.50)} = 1.36 \end{aligned} \tag{28}$$

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, and year estimates.

For example, if the grade 8 Math and grade 8 ELA classes had no students in common, then their correlation would be zero. On the other hand, if grade 8 Math and grade 8 ELA classes contained many of the same students, there would be a positive correlation. However, even if those two classes had exactly the same students, the correlation would likely be considerably less than +1. Correlations of gains across years or subjects might be positive or slightly negative as the same student's score can be used in multiple gains. The actual correlations and covariances themselves are obtained as part of the modeling process using equation (10) from Section 2.2.4.1. It would be impossible to obtain them outside of the modeling process. This process uses all the information about which students are in which subject and grade 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 will typically be between these two extremes, which are 0.97 and 1.36 in this 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 those students have will contribute to the magnitude of the standard error.

5.1.2.3 Composite Index Across Subjects

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

$$Composite\ Index = \frac{Composite\ Gain}{Composite\ Standard\ Error} = \frac{1.83}{1.15} = 1.59 \tag{29}$$

Although some of the values in the example were rounded for display purposes, the actual rounding or truncating occurs only after all the measures have been combined as described in Section 4.5.

5.1.3 Technical Description of the Composite Index Based on Predictive Model Measures

For our sample teacher (and for the majority of teachers who receive reporting from the predictive model in Ohio), there is only one available value-added measure from the predictive model. This means that the reported value-added index for that subject will be the same that is calculated for the predictive-based composite index. For the sample teacher, only an Algebra I growth measure is available.

$$Composite\ Index = \frac{Alg\ I\ Growth\ Measure}{Alg\ I\ Standard\ Error} = \frac{11.75}{6.20} = 1.90 \tag{30}$$

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

5.1.4 Technical Description of the Combined Composite Index Across Subjects Based on the Gain and Predictive Models

The two composite indices from the gain and predictive models are 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 gain model and 20 to the predictive model. The combined composite index would be calculated as follows using these weightings, the gain-based composite index across subjects, and the predictive-based index across subjects:

$$Unadjusted\ Combined\ Comp\ Index = \left(\frac{135}{155}\right)(1.59) + \left(\frac{20}{155}\right)(1.90) = 1.62 \tag{31}$$

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

$$Final\ Combined\ Comp\ SE = \sqrt{\left(\frac{135}{155}\right)^2 (1)^2 + \left(\frac{20}{155}\right)^2 (1)^2} = 0.88 \tag{32}$$

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

5.2 District and School Composites

5.2.1 Overview

This section presents how school-level and district-level composites are calculated. The key policy decisions for school and district composites can be summarized as follows:

- For the 2022-23 reporting, school and district composites displayed on accountability reports will be calculated for 2022-23 and for an up-to-two years composite that includes data from both 2021-22 (weighted at 33%) and 2022-23 (weighted at 67%).
- The up-to-two years composites will only be calculated for schools or districts with 2022-23 results.
- School and district accountability composites use both an effect size and index component.
- The composites for districts and schools can include OST Mathematics, ELA, Science, Algebra I, Mathematics I, Geometry, Mathematics II, ELA II, Biology, American US History, and American US Government.
- The single-year composites for schools and districts weight each subject and grade by the number of scores in that subject and grade.

The key steps for determining a school or district’s single-year composite effect size and index are as follows:

1. For growth measures based on the gain model, calculate the effect size and effect size standard error across subjects and grades.
2. For growth measures based on the predictive model, create an effect size and effect size standard error for each subject.
3. Combine the gain model composite effect size and standard error values with the predictive model effect sizes and standard errors assuming independence. Create the index value by dividing the resulting effect size standard error.

The following sections illustrate this process for a single-year composite using value-added measures from a sample middle school, which are provided below.

Table 6: Sample School Value-Added Information

Year	Subject	Grade	Value-Added Gain	Standard Error	Standard Deviation	Effect Size	Standard Error of Effect Size	Number of Students
1	Math	6	3.30	0.70	12.25	0.27	0.06	44
1	ELA	6	-1.10	1.00	13.36	-0.08	0.07	46
1	Math	7	2.00	0.50	11.9	0.17	0.04	50
1	ELA	7	2.40	1.10	13.32	0.18	0.08	50
1	Math	8	-0.30	0.60	12.45	-0.02	0.05	40
1	ELA	8	3.80	0.70	12.31	0.31	0.06	50
1	Algebra I	N/A	-11.50	6.20	17.7	-0.65	0.35	35

5.2.2 Calculation of the Composite Effect Size and Index based on Gain Model Measures

In the past, the district or school composite index for the gain model was calculated by dividing the composite gain by its composite standard error as described in the teacher composite section. Beginning with the 2021-22 reporting, the district or school composite index for the gain model will be calculated by dividing the composite effect size by the composite standard error of the effect size since both the effect size and index components are needed. The calculations for each of these metrics are provided below.

5.2.2.1 Composite Effect Size Across Subjects

Effect sizes are already standardized to the same scale, so the composite effect size across the six subjects and grades from the gain model is a weighted average based on the number of scores in each subject and grade. For the school, the total number of students affiliated with growth measures from the gain model is 44 + 46 + 50 + 50 + 40 + 50, or 280 students. The Math grade 6 growth measure would be weighted at 44/280, the ELA grade 6 growth measure would be weighted at 46/280, and so on.

The calculation of the composite effect size across subjects based on the gain model is as follows:

$$\begin{aligned} Composite\ Effect\ Size &= \frac{44}{280}Math_6 + \frac{46}{280}ELA_6 + \frac{50}{280}Math_7 + \frac{50}{280}ELA_7 + \frac{40}{280}Math_8 + \frac{50}{280}ELA_8 \\ &= \left(\frac{44}{280}\right)(0.27) + \left(\frac{46}{280}\right)(-0.08) + \left(\frac{50}{280}\right)(0.17) + \left(\frac{50}{280}\right)(0.18) + \left(\frac{40}{280}\right)(-0.02) + \left(\frac{50}{280}\right)(0.31) = 0.14 \end{aligned} \tag{33}$$

5.2.2.2 Composite Standard Error of Effect Size Across Subjects

Similar to the teacher example, the standard error of the district or school composite effect size based on the gain model cannot be calculated using the assumption that the effect sizes making up the composite are independent. This is because many of the same students are likely represented in different year/subject/grades, such as grade 8 Math and grade 8 ELA. The statistical approach, outlined in Section 2.2.4 (with references), is quite sophisticated and will account for the correlations between pairs of effect sizes as shown in equation (25) and using equation (6) for schools and equation (10) for teachers.[5] The composites are indeed linear combinations of the fixed effects of the models and can be estimated as described in Section 2.2.4. The magnitude of each correlation depends on the extent to which the same students are in both estimates for any two subject, grade, and year estimates.

To demonstrate the impact of covariance terms on the standard error of the effect size, it is useful to calculate the standard error of the effect size using (again, inappropriately) the assumption of independence. Using the student weightings and standard errors reported in Table 5 and assuming total independence, the standard error would then be as follows, where "SE" represents the standard error of the composite effect size:

$$\begin{aligned} Composite\ SE &= \sqrt{\begin{aligned}&\left(\frac{44}{280}\right)^2 (SE\ Math_6)^2 + \left(\frac{46}{280}\right)^2 (SE\ Read_6)^2 + \left(\frac{50}{280}\right)^2 (SE\ Math_7)^2 \\ &+ \left(\frac{50}{280}\right)^2 (SE\ Read_7)^2 + \left(\frac{40}{280}\right)^2 (SE\ Math_8)^2 + \left(\frac{50}{280}\right)^2 (SE\ Read_8)^2\end{aligned}} \\ &= \sqrt{\begin{aligned}&\left(\frac{44}{280}\right)^2 (0.06)^2 + \left(\frac{46}{280}\right)^2 (0.07)^2 + \left(\frac{50}{280}\right)^2 (0.04)^2 \\ &+ \left(\frac{50}{280}\right)^2 (0.08)^2 + \left(\frac{40}{280}\right)^2 (0.05)^2 + \left(\frac{50}{280}\right)^2 (0.06)^2\end{aligned}} = 0.03 \end{aligned} \tag{34}$$

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

The actual standard error will likely be above the value of 0.03 due to students being in both Math and ELA in the school with the specific value depending on the values of the correlations between pairs of value-added gains. The magnitude of each correlation depends on the extent to which the same students are in both estimates for any two subject, grade, and year estimates.

For the sake of simplicity, let us assume that the actual standard error of the effect size was 0.05 for the school composite standard error in this example.

5.2.2.3 Composite Index Across Subjects

The next step is to calculate the school composite index based on the gain model, which is the composite effect size divided by the composite standard error of the effect size. The gain-based composite index for this school would be calculated as follows:

$$Composite\ Index = \frac{Composite\ Effect\ Size}{Composite\ Standard\ Error\ of\ the\ Effect\ Size} = \frac{0.14}{0.05} = 2.80 \tag{35}$$

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

Although some of the values in the example were rounded for display purposes, the actual rounding or truncating occurs only after all the measures have been combined as described in Section 4.5.

5.2.3 Calculation of the Effect Size and Index Based on Predictive Model Measures

For our sample school (and for the majority of middle schools in Ohio), there is only one available growth measure from the predictive model, which simplifies our example. For the sample school, the index based on the predictive model is calculated as follows.

$$Index = \frac{Alg\ I\ Growth\ Measure}{Alg\ I\ Standard\ Error} = \frac{-11.50}{6.20} = -1.85 \tag{36}$$

This equation could also be expressed as the effect size divided by the standard error of the effect size, which would yield the same result. The effect size for this growth measure is -0.65 as denoted in Table 5 above.

Should a school or district have more than one value-added measure based on the predictive model, then multiple effect size and indices would be incorporated into the next step in the calculation.

5.2.4 Calculation of the Combined Composite Index Across Subjects Based on the Gain and Predictive Models

The effect size and index from the gain model is combined with the effect sizes and indices from the predictive models weighting each according to the number of students' scores within each model to determine the combined composite effect size and index.

Our sample school has 315 students of which 280 are linked to the gain model and 35 to the predictive model for Algebra I. The combined composite effect size would be calculated as follows using these weightings, the gain-based composite effect size across subjects, and the predictive-based effect sizes:

$$Composite\ Combined\ Effect\ Size = \left(\frac{280}{315}\right)0.14 + \left(\frac{35}{315}\right)(-0.65) = 0.05 \tag{37}$$

The combined composite index is calculated by creating a standard error of this combined effect size and then dividing the combined effect size by the standard error of that effect size:

$$Combined\ Effect\ Size\ SE = \sqrt{\left(\frac{280}{315}\right)^2 (0.05)^2 + \left(\frac{35}{315}\right)^2 (0.35)^2} = 0.06 \tag{38}$$

$$Combined\ Composite\ Index = \frac{0.05}{0.06} = 0.83 \tag{39}$$

Therefore, the final combined composite index value is 0.05 divided by 0.06, or 0.83. Different accountability measures use different subsets of students, but the approach to calculating the overall composite is the same.

5.2.5 Calculation of the Up-to-Two Years Composite

The up-to-two years composite uses the values described in Section 5.2.4 from each year to calculate the same outputs across the two applicable years. These composites use a weighted average for each of the two years, with the results from 2022-23 weighted at 67 percent and the results from 2021-22 weighted at 33 percent. This means that instead of the number of students used in equation (37), these weightings for each of the years are used. After the composite combined effect size is calculated, the combined effect size standard error is calculated using the same approach outlined in equation (38). Finally, as in equation (39), the composite combined effect size is divided by the combined effect size standard error.

5.2.6 Categorizing Accountability Composites

Beginning with the analysis using data from the 2021-22 school year, accountability composites use both the index, which is calculated using the rules outlined earlier in this section, and the effect size described in Section 4.4.

To calculate the effect size for the overall composite, each growth measure is divided by the student-level standard deviation of growth. This value is a constant within each year subject and grade. The composite effect size is a weighted average of the effect sizes based on the FTE number of students. This weighted average is calculated like the weighted average described in Section 5.2.2.1, with the one exception being that both the gain and predictive model effect sizes are included in the weighted average.

After the effect size and growth index are calculated, accountability composites categorize growth measures using a two-step process.

The growth index is the growth estimate divided by the standard error, which is specific to each estimate. The effect size is the growth measure divided by the student-level standard deviation of growth. The effect size provides an indicator of magnitude and practical significance that the group of students met, exceeded, or fell short of expected growth.

This two-step approach first considers whether there is statistical certainty that the growth measure is above or below the expectation of growth. The second step determines whether the growth measure is above or below the growth expectation by a certain magnitude. The first step uses the growth index to determine thresholds for the certainty, and the second step uses the effect size to determine thresholds for magnitude.

For the first step with uncertainty, the thresholds are an index of +2 or greater, an index of -2 or less, or an index between -2 and +2. These thresholds are similar to the concept of a 95% confidence interval. If a 95% confidence interval around the growth measure did not contain the growth expectation, then they would fall outside the thresholds. The second step uses an effect size threshold of 0.1 and -0.1 for districts and 0.2 and -0.2 for schools.

These two steps are used to assign five categories. The top category, Five Stars, has a growth index of greater than or equal to 2 *and* an effect size of greater than or equal to 0.1 (for districts) or 0.2 (for schools). When the index is greater than or equal to 2 but the effect size is less than 0.1 (for districts) or 0.2 (for schools), the measure is categorized as Four Stars. The next highest category, Three Stars, includes all measures where the growth index is less than 2 and greater than or equal to -2. When the growth index is less than -2, measures are categorized as Two Stars when the effect size is greater than or equal to -0.1 (for districts) or -0.2 (for schools) and are categorized as One Star when the effect size is less than -0.1 (for districts) or -0.2 (for schools).

The table below provides the color-coding, definitions, and interpretation for accountability composites. These categories are used for both the single-year and up-to-two years composites.

Table 7: Accountability Categories, Definitions, and Interpretations

Category	Definition	Interpretation
Five Stars	Index is greater than or equal to 2 *and* the effect size is greater than or equal to 0.1 (districts) or 0.2 (schools)	Significant evidence that the school or district exceeded growth expectations by a larger magnitude
Four Stars	Index is greater than or equal to 2 *and* the effect size is less than 0.1 (districts) or 0.2 (schools)	Significant evidence that the school or district exceeded student growth expectations
Three Stars	Index is less than 2 and greater than or equal to -2	Evidence that the school or district met student growth expectations
Two Stars	Index is less than -2 *and* the effect size is greater than or equal to -0.1 (districts) or -0.2 (schools).	Significant evidence that the school or district fell short of student growth expectations
One Star	Index is less than -2 *and* the effect size is less than -0.1 (districts) or -0.2 (schools).	Significant evidence that the school or district fell short of student growth expectations by a larger magnitude.

NOTE: When an index or effect size falls exactly on the boundary between two categories, the higher category is assigned.

5.3 Principal Composites

5.3.1 Overview

Principal composites use school growth measures, and the key policy decisions for principal composites can be summarized below.

- The term “principal” here refers to both assistant principals and principals. They are equivalent for the purposes of the calculations, and composites should be calculated for each person rather than per person per position.
- There are principals who fill the role of principal (P) for more than one school at a time.
- Many schools have more than one assistant principal (AP) at a time.
- Assume that each school has a single principal at any given time until, after applying the business rules below, the data still show more than one principal at a school in a particular school year.
- Schools named “district testing” will be excluded.
- A principal (P) or assistant principal (AP) must be in the school for 120 school days (190 calendar days) of a single school year to be linked to a school for that year. If someone was both a principal and an assistant principal, the days are combined to reflect the total time in both positions before determining whether this requirement is met.

- The cutoff of a school year that is used for examining the data is from 9/15 to 5/31.
- If a principal or assistant principal starts in a school after 5/31 and ends the position before 9/15, do not link the staff member to that school for that year.

5.3.2 Multiple Principals Reported at a School Per Year

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

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

5.3.3 Calculation

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

- Calculate a composite for each person. Treat assistant principal and principal positions as equivalent.
- In order to receive a composite, a principal must be assigned to a school with a growth measure in the most recent year to receive a composite. If the principal is not assigned to a school in the most recent year or the school to which they are assigned does not have a composite based on growth measures, then the principal will not receive a composite for that year.
- If a principal remains in the same school within a year, calculate the principal's single-year estimate as the school composite across subjects and grades for that year.
- If a principal was in different schools within a year, calculate the principal's single-year estimate (across schools) as the weighted average, adjusted for standard error, of the school composites for that year assuming independence. The weights are based on the number of subjects/grades in each school for that year.

6 Input Data Used in the Ohio Growth Model

6.1 Assessment Data Used in Ohio

For the analysis and reporting based on the 2022-23 school year, EVAAS receives the following assessments for use in the growth and/or projection models:

- OST Mathematics in grades 3–8
- OST English Language Arts (ELA) in grades 3–8
- OST Science in grades 5 and 8
- OST Algebra I, Mathematics I and II, and Geometry
- OST English Language Arts (ELA) II
- OST Biology
- OST American US History and American US Government
- ACT Mathematics, English, and Reading
- SAT Evidence-Based Reading and Writing, and Mathematics
- AP Biology
- AP Calculus AB
- AP Chemistry
- AP Computer Science
- AP English Language and Composition
- AP English Literature and Composition
- AP Environmental Science
- AP European History
- AP Human Geography
- AP Macroeconomics
- AP Microeconomics
- AP Physics
- AP Psychology
- AP Statistics
- AP U.S. Government and Politics
- AP U.S. History
- AP World History

OSTs are administered in the spring semester of the school year, except for ELA in grade 3 and the EOCs, which are administered during the fall, spring, and summer semesters. In grade 3, the highest of the two scores for each student is used in the growth models, which is consistent with the accountability rules in Ohio.

The ACT or SAT assessment is administered to all students across the state in the spring of grade 11.

In the past, some districts also received value-added reporting based on extended testing (vendor assessments that measure subjects and grades outside the state testing scope), but this reporting is not available because these assessments have not been administered since the 2017-18 school year.

Assessment files provide the following data for each student score:

- Scale score
- Performance level
- Test taken
- Tested grade
- Accountable district IRN
- Accountable org IRN
- Testing district IRN
- Testing org IRN
- Reporting district IRN
- Reporting org IRN

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

6.2 Student Information

Ohio’s state law prohibits ODE from maintaining student names; therefore, ODE provides assessment files that include only the state student ID (SSID) for each student and no name information. IBM contracts with the State of Ohio to maintain the crosswalk with student names and IDs, so IBM securely transfers student names to Battelle for Kids (BFK) and The Management Council of the Ohio Education Computer Network (MCOECN). Those student names are matched using SSID and sent to EVAAS. The file from IBM contains the following:

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

The secure EVAAS website includes student names in reports restricted to authorized users in Local Education Agencies (LEAs) for further analysis and improvement purposes.

This student information (not names) is also used for accountability categories that are reported to the public. EVAAS receives various socioeconomic, demographic, and programmatic identifiers in the student data system. Currently, these categories are:

- Gifted
- Gifted – ELA
- Gifted – Math
- Gifted – Science
- Gifted – Superior Cognitive
- Migrant
- English Learner
- Economically Disadvantaged
- Students with Disabilities
- Gender
- Race
  - American Indian
  - Asian/Pacific Islander (This includes Asian and Hawaiian/Other Pacific Islander)

  - Black
  - Hispanic
  - White
  - Multi-Racial
- Chronically Absent

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

6.3 Teacher Information

In order to provide accurate and verified student-teacher linkages in the teacher growth models, Ohio educators are given the opportunity to complete roster verification. This process enables teachers to confirm their class rosters for students in a particular subject, grade, and year, and it captures scenarios where multiple teachers have instructional responsibility for students. Administrators also verify the linkages as an additional check. Roster verification, therefore, increases the reliability and accuracy of teacher-level analyses.

SAS receives teacher identification data and student-teacher linkages from MCOECN. The student-teacher linkage files include the following information:

- District IRN
- District name
- School IRN
- School name
- Teacher state ID
- Teacher name
- Teacher email
- Student identifying information, including SSID
- Student grade level
- Subject of instruction
- Percentage claimed by teacher

More information about teacher roster verification is available at the following website: http://education.ohio.gov/Topics/Teaching/Educator-Evaluation-System/Ohio-s-Teacher-Evaluation-System/Student-Growth-Measures/Value-Added-Student-Growth-Measure/Value-Added-Roster-Verification.

If districts do not participate in roster verification, then the teacher-student linkage reported and verified through EMIS is used in the teacher growth models. The EMIS teacher-student linkage files contain similar information to that provided in the roster verification teacher-student linkage files. However, teacher claiming percentages are not included.

If a teacher in a district that participates in roster verification has claimed a student in a course/subject/grade, and that same student was reported for that same course/subject/grade in one of the districts relying on EMIS teacher-student linkages, the percentage responsibility claimed in the roster verification linkage is used primarily, and then the EMIS linkage data is used for any remaining instructional responsibility, up to 100%.

6.4 Principal Information

EVAAS receives two data files from ODE on individual principals and assistant principals. One file provides a list of principals and assistant principals in every school, their employment information, and their position start and end dates for that position as reported into EMIS by districts, community schools, Joint Vocational School Districts, and Educational Service Centers. The other file provides a listing of principals and assistant principals in every school, their employment information, and the school year in which they are reported in that position by districts, community schools, Joint Vocational School Districts, and Education Service Centers.

For the purposes of principal composites in Section 5.3, both assistant principals and principals are considered the same.

7 Business Rules

7.1 Assessment Verification for Use in Growth Models

To be used appropriately in any growth models, the scales of these assessments must meet three criteria:

1. **There is sufficient stretch in the scales** to ensure progress can be measured for both low-achieving students as well as high-achieving students. A floor or ceiling in the scales could disadvantage educators serving either low-achieving or high-achieving students.
2. **The test is highly related to the academic standards** so that it is possible to measure progress with the assessment in that subject, grade, and year.
3. **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, and year. This will be monitored each subsequent year that the test is given.

These criteria are checked annually for each assessment prior to use in any growth model, and Ohio's current standardized assessments meet them. These criteria are explained in more detail below.

7.1.1 Stretch

Stretch indicates whether the scaling of the assessment permits student growth to be measured for both very low- or very high-achieving students. A test "ceiling" or "floor" inhibits the ability to assess students' growth for students who would have otherwise scored higher or lower than the test allowed. It is also important that there are enough test scores at the high or low end of achievement, so that measurable differences can be observed.

Stretch can be determined by the percentage of students who score near the minimum or the maximum level for each assessment. If a much larger percentage of students scored at the maximum in one grade than in the prior grade, then it might seem that these students had negative growth at the very top of the scale when it is likely due to the artificial ceiling of the assessment. Percentages for all OST assessments are well below acceptable values, meaning that the OSTs have adequate stretch to measure value-added even in situations where the group of students are very high or low achieving.

In 2023, the percentage of students who achieved a maximum score on the OST end-of-grade and end-of-course assessments ranged from a high of 2.75% (fourth-grade Math) to a low of .01% (ELA II).

7.1.2 Relevance

Relevance indicates whether the test is sufficiently aligned with the curriculum. The requirement that tested material correlates with standards will be met if the assessments are designed to assess what students are expected to know and be able to do at each grade level. Since the OSTs are designed to measure state curriculum, this criterion is met by the OSTs.

7.1.3 Reliability

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

7.2 Pre-Analytic Processing

7.2.1 Missing Grade

In Ohio, the grade used in the analyses and reporting is the tested grade, not the enrolled grade. If a grade is missing on an end-of-grade test record, then that record will be excluded from all analyses. The grade is required to include a student's score in the appropriate part of the models and to convert the student's score into the appropriate NCE in the gain-based model.

Of the 1,790,617 records from the 2022-23 OST Mathematics, ELA, and Science assessments, no records were excluded due to this business rule.

7.2.2 Duplicate (Same) Scores

If a student has a duplicate score for a particular subject and tested grade in a given testing period and the duplicate score is exactly the same in both records, then the following business rules will be applied in order:

- If there are multiple records for a student and one record has an accountable district but the other records do not have an accountable district, then the analysis will use the record that contains the accountable district.
- If the student has multiple records that are identical except that one has an accountable school and the other does not, then the analysis will keep the record that contains the accountable school.
- If the student is now accountable to multiple districts/schools, both records will be excluded**.**

If there are still multiple records for a student at this point, then the above conditions are not met, but EVAAS can now apply similar rules at the tested level:

- If the student has multiple records and one record contains a tested district but the tested district is missing in the other record, then the analysis will keep the record that contains the tested district.
- If the student has identical records except that there is tested school information in one record and no tested school information in the other record, then the analysis will keep the record that contains the tested school information. If there are still multiple records for the student, this means that those records have the same accountable and tested information. If one record is linked to a teacher and the other is not, then the analysis will keep the record that is linked to a teacher.
- If there are still multiple records, then the analysis will keep the record that has the most demographic fields filled out.
- If there are still multiple records, then the analysis will only keep one record.

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

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

Of the 2,572,344 records from the 2022-23 OST Mathematics, ELA, Science, Algebra I, Geometry, ELA II, Mathematics I, Mathematics II, Biology, American US History, and American US Government assessments, 1,934 records (0.08%) were excluded due to this business rule.

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

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

Of the 2,572,344 records from the 2022-23 OST Mathematics, ELA, Science, Algebra I, Geometry, ELA II, Mathematics I, Mathematics II, Biology, American US History, and American US Government assessments, 2,052 records (0.08%) were excluded due to this business rule.

7.2.5 Students with Multiple Grade Levels in the Same Subject in the Same Year

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

Of the 2,572,344 records from the 2021-22 OST Mathematics, ELA, Science, Algebra I, Geometry, ELA II, Mathematics I, Mathematics II, Biology, American US History, and American US Government assessments, 6 records (0.0002%) were excluded due to this business rule.

7.2.6 Students with Records That Have Unexpected Grade Level Changes

If a student skips more than one grade level (e.g., moves from sixth in 2022 to ninth in 2023) or is moved back by one grade or more (i.e., moves from fourth in 2022 to third in 2023) 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. Per ODE's decision, the analysis does not remove students with scores that appear to be associated with inconsistent grades. The analysis leaves students in the analysis at the tested grade that EVAAS receives from ODE.

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

If a student is tested at two different accountable schools in a given testing period, then the student's records are examined to determine whether two separate students were inadvertently combined. If this is the case, then the student data is adjusted so that each unique student is associated with only the appropriate scores. In Ohio, it can happen if a student is accelerated in a subject and tests at two different accountable schools.

7.2.8 Outliers

Student assessment scores are checked each year to determine whether they are outliers in context with all the other scores in a reference group of scores from the individual student. These reference scores are weighted differently depending on proximity in time to the score in question. Scores are checked for outliers using related subjects as the reference group. For example, when searching for outliers for Math test scores, all OST Math grades are examined simultaneously, and any scores that appear inconsistent, given the other scores for the student, are flagged. Scores are flagged in a conservative way to avoid excluding any student scores that should not be excluded. Scores can be flagged as either high or low outliers. Once an outlier is discovered, that outlier will not be used in the analysis, but it will be displayed on the student testing history on the 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. Using this t-value, the growth models 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 Math and ELA in grades 3–8) or -4.0 (for other assessments) when looking at the difference between the score in question and the reference group of scores.
- The percentile of the comparison score must be above a certain value. This value depends on the position of the individual score in question but will range from 10 to 90 with the ranges of the individual percentile score.

For high-end outliers, the rules are:

- The percentile of the score must be above 50.
- The t-value must be above 4.5 (for Math and ELA in grades 3–8) or 5.0 (for other assessments).
- The percentile of the comparison score must be below a certain value.
- There must be at least three scores in the comparison score average.

Of the 2,572,344 records from the 2022-23 OST Mathematics, ELA, Science, Algebra I, Geometry, ELA II, Mathematics I, Mathematics II, Biology, American US History, and American US Government assessments, 787 records (0.03%) were excluded due to this business rule.

7.2.9 Linking Records over Time

Each year, EVAAS receives data files that include student assessment data and file formats. These data are checked each year prior to incorporation 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 the student identification numbers provided by the state. Teacher records are matched over time using the teacher credential ID only as requested by ODE because other information, such as teacher name, might change over time, but the credential ID remains the same.

7.3 Growth Models

7.3.1 Students Included in the Analysis

As described in Section 7.2, student scores might be excluded due to the business rules, such as outlier scores.

For the gain model, all students are included in these analyses if they have assessment scores that can be used. The gain model uses all available OST Mathematics and ELA results for each student. Because this model follows students from one grade to the next and measures growth as the change in achievement from one grade to the next, the gain model assumes typical grade patterns for students. Students with non-traditional patterns, such as those who have been retained in a grade or skipped a grade, are treated as separate students in the model. In other words, these students are still included in the gain model, but the students are treated as separate students in different cohorts when these non-traditional patterns occur. This process occurs separately by subject since some students can be accelerated in one subject and not in another.

For the predictive and projection models, a student must have at least three valid predictor scores that can be used in the analysis, all of which cannot be deemed outliers. (See Section 7.2.8 on Outliers.) These scores can be from any year, subject, and grade that are used in the analysis. In other words, the student's expected score can incorporate other subjects beyond the subject of the assessment being used to measure growth. The required three predictor scores are needed to sufficiently dampen the error of measurement in the tests to provide a reliable measure. If a student does not meet the three-score minimum, then that student is excluded from the analyses. It is important to note that not all students have to have the same three prior test scores; they only have to have some subset of three that were used in the analysis. Unlike the gain model, students with non-traditional grade patterns are included in the predictive model as one student. Since the predictive model does not determine growth based on consecutive grade movement on tests, students do not need to stay in one cohort from one year to the next. That said, if a student is retained and retakes the same test, then that prior score on the same test will not be used as a predictor for the same test as a response in the predictive model. This is mainly due to the fact that very few students used in the models have a prior score on the same test that could be used as a predictor. In fact, in the predictive model, it is typically the case that a prior test is only considered a possible predictor when at least 50% of the students used in that model have those prior test scores.

In the teacher analysis for both the gain and predictive models, students are excluded if they have more absences than an amount originally prescribed by state law and now set by ODE policy, which is currently 45 excused or unexcused days. ODE provides EVAAS with a file that flags students who should be excluded based on that policy.

7.3.2 Minimum Number of Students to Receive a Report

The growth models require a minimum number of students in the analysis in order for districts, schools, and teachers to receive a growth report. This is to ensure reliable results.

7.3.2.1 District and School Model

For the gain model, the minimum student count to report an estimated average NCE *score* (i.e., either entering or exiting achievement) is six students in a specific subject, grade, and year. To report an estimated NCE *gain* in a specific subject, grade, and year, there are additional requirements:

- There must be at least six students who are associated with the school or district in the subject, grade, and year. For the accountable analysis, this minimum must be met by accountable students. For the tested analysis, this minimum must be met by tested students.
- Of those students who are associated with the school or district in the current year and grade, there must be at least six students in each subject, grade, and year in order for that subject, grade, and year to be used in the gain calculation.
- 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 and grade as well as the prior year and grade in the same subject. However, due to the rule above, it is typically the case that at least six students have a “simple gain.” In some cases where students only have a Math or ELA score in the current year or previous year, this value dips below six.
- For any district or school growth measures based on specific student groups, the same requirements described above apply for the students in that specific student group.

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

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

For the predictive model, the minimum student count to receive a growth measure is 10 students in a specific subject, grade, and year. These students must have the required three prior test scores needed to receive an expected score in that subject, grade, and year.

7.3.2.2 Teacher Model

The teacher *gain model* includes teachers who are linked to at least six students with a valid test score in the same subject, grade, and year. This requirement does not consider the percentage of instructional time that the teacher spends with each student in a specific subject and grade.

The teacher *predictive model* includes teachers who are linked to at least 10 students with a valid test score in the same subject/grade or course within a year. This requirement does not consider the percentage of instructional time the teacher spends with each student in a specific subject and grade.

For both the gain and predictive models, to receive a Teacher *report* in a particular year, subject, and grade, there is an additional requirement. A teacher must have at least six Full Year Equivalent (FYE) students in a specific subject, grade, and year. The teacher’s number of FYE students is based on the number of students linked to that teacher and the percentage of instructional time the teacher has for each student. For example, if a teacher taught 10 students for 50% of their instructional time, then the teacher’s FYE number of students would be five, and the teacher would not receive a teacher growth report. If another teacher taught 12 students for 50% of their instructional time, then that teacher would have six FYE students and would receive a teacher growth report. The instructional time attribution is obtained from the linkage roster verification process that is used in Ohio.

The teacher gain model has an additional requirement. The teacher must be linked to at least five students with prior test score data in the same subject, and the test data can come from any prior grade as long as they are part of the student's regular cohort. One of these five students must have a "gain," meaning the same subject prior test score must come from the immediate prior year and prior grade. Note that if a student repeats a grade, then the prior test data would not apply as the student has started a new cohort.

7.3.3 Accountable and Tested Analyses

Many assessments include two sets of reports: one for students who are considered accountable to the district or school and another for students who tested at the district or school. The definitions of accountable and tested are based on the business rules governing the accountability system. More information about the "Full Academic Year/Where Kids Count Rules" is available at the following link: https://education.ohio.gov/Topics/Data/Report-Card-Resources/Resources-and-Technical-Document.

In most cases, the "accountable" district and school are the same as the "tested" district and school. However, there are some cases where these are different. For example, there could be students with disabilities who are held accountable to a different school or only the district level and not the school where they might have tested. There are also students who are accountable to the district or the state for various purposes.

7.3.4 Accountable Student Groups

7.3.4.1 Overall Measures for Districts and Schools

The overall analysis includes only those students who are included in the accountable student set. It includes all students who are accountable to a district or school.

7.3.4.2 Gifted Students for Districts and Schools

The gifted student analysis includes only those students who are included in the accountable student set. Students are included in the Math analysis if they are either identified as gifted in Math or superior cognitive. In the Math analysis, students' prior and current Math and ELA test scores are included. Similarly, for Reading, students are included who are identified as gifted in ELA or superior cognitive. All other Math and ELA scores from those students are included in the ELA analysis.

7.3.4.3 Students with Disabilities for Districts and Schools

The students with disabilities analysis includes only those students who are included in the accountable student. Students are included in the analysis if they are denoted as students with disabilities as recorded by the disability flag in EMIS.

7.3.4.4 ESSA Accountability Student Groups for Districts and Schools

Ohio uses student group growth measures in its federal accountability system. The student groups include White, Black, Asian/Pacific Islander, American Indian, Multi-Racial, Hispanic, EL, and ED. These measures are provided using the OST subjects with a composite across Math in grades 4–8 and ELA in grades 4–8. As with the regular model, student group value-added measures require six students in each subject and grade to be created.

7.3.4.5 Community School Closure

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

7.4 Student-Teacher Linkages

Student-teacher linkages are not used in the analysis if the students are listed as having more than 45 unexcused absences.

Of the 2,769,427 linkages from the 2022-23 OST Mathematics, ELA, Science, Algebra I, Geometry, ELA II, Mathematics I, Mathematics II, Biology, American US History, and American US Government assessments, 110,869 linkages (4%) were excluded due to this business rule.

Student-teacher linkages are connected to assessment data based on the subject and identification information described in Section 6.3. The model will make adjustments to linkages if a student is claimed by teachers at a total percentage higher than 100% in an individual year, subject, and grade. If over-claiming happens, then the individual teacher’s weight is divided by the total sum of all weights to redistribute the attribution of the student's test scores across teachers. Underclaimed linkages for students are not adjusted because a student can be claimed less than 100% for various reasons (such as a student who lives out of state for part of the year).