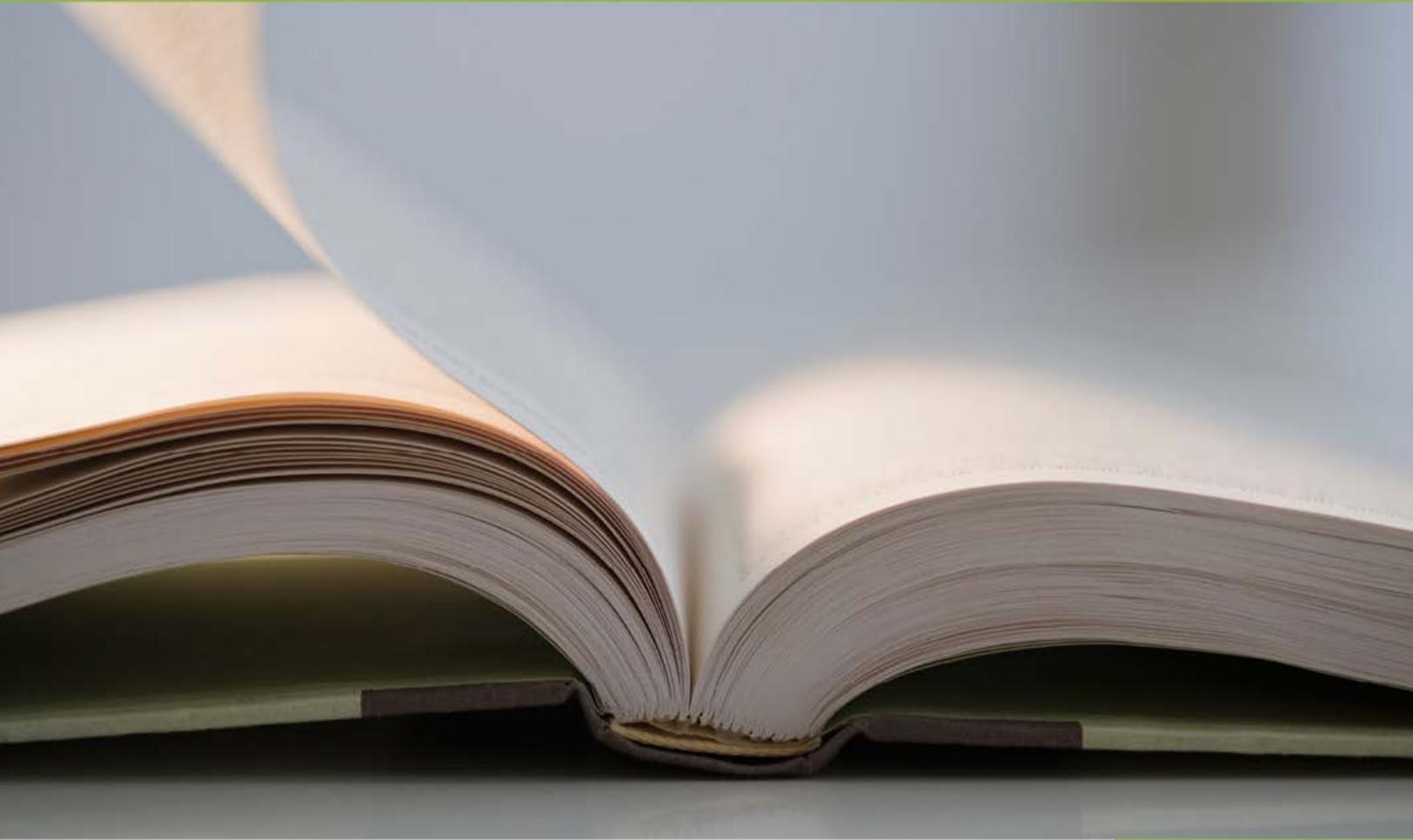


SAS® EVAAS®

Technical Documentation of PVAAS Analyses



Contents

1	Introduction to value-added reporting in Pennsylvania	1
2	Input data used in PVAAS	2
2.1	Determining suitability of assessments	2
2.1.1	Current assessments	2
2.1.2	Transitioning to future assessments	2
2.2	Assessment data used in Pennsylvania	2
2.2.1	Tests given in consecutive grades for the same subject	3
2.2.1	Tests given in non-consecutive grades for the same subject	3
2.2.2	Student identification information	3
2.2.3	Assessment information provided	3
2.2.4	General exclusion of non-public school assessment data	3
2.3	Student-level information	3
2.4	Teacher-level information	4
3	Value-added analyses	6
3.1	Multivariate Response Model (MRM) reporting for tests in consecutive grades	7
3.1.1	MRM at the conceptual level	8
3.1.2	Normal curve equivalents	9
3.1.3	Standard statistical notation of the linear mixed model and the MRM	11
3.1.4	Where the MRM is used in Pennsylvania	16
3.1.5	Students included in the analysis	16
3.1.6	Minimum number of students for reporting	17
3.2	Univariate Response Model (URM) for tests in non-consecutive grades	18
3.2.1	URM at the conceptual level	19
3.2.2	Standard statistical notation of the district, school and teacher models	19
3.2.3	Students included in the analysis	21
3.2.4	Minimum number of students for reporting	22
3.2.5	Prior tests used in the URM analyses	22
4	Growth expectation	23
4.1	Intra-year approach	23
4.1.1	Description	23
4.1.2	Illustrated example	24
4.2	Base year approach	24
4.2.1	Description	24
4.2.2	Illustrated example	25
4.3	Defining the expectation of growth during an assessment change	26
5	Using standard errors to create levels of certainty and define effectiveness	27
5.1	Using standard errors derived from the models	27
5.2	Defining evidence of growth in terms of standard errors	27
5.3	Rounding and truncating rules	28
5.4	Other scales used for reporting in Pennsylvania	28
6	Teacher multi-year composite calculation	30
6.1	Overview of teacher level composites	30
6.2	Calculating the index	30

6.3	Combining the index values across subjects, grades and years	31
7	PVAAS Projection Model	32
8	Data quality and pre-analytic data processing.....	34
8.1	Data quality.....	34
8.2	Checks of scaled score distributions.....	34
8.2.1	Stretch.....	34
8.2.2	Relevance.....	34
8.2.3	Reliability.....	34
8.3	Data quality business rules	35
8.3.1	Missing grade levels.....	35
8.3.2	Duplicate (same) scores.....	35
8.3.3	Students with missing districts or schools for some scores but not others	35
8.3.4	Students with multiple (different) scores in the same testing administration.....	35
8.3.5	Students with multiple grade levels in the same subject in the same year.....	35
8.3.6	Students with records that have unexpected grade level changes	36
8.3.7	Students with records at multiple schools in the same test period	36
8.3.8	Outliers.....	36

1 Introduction to value-added reporting in Pennsylvania

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

- Growth = current achievement/current results compared to all prior achievement/prior results, with achievement being measured by a quality assessment such as the PSSA and Keystone tests.

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

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

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

The State of Pennsylvania and the EVAAS team have provided value-added reporting since 2003. The initial collaboration began with a pilot group of districts, and this expanded to statewide district and school value-added reporting by 2006. In 2014, teacher value-added reports also became available for the state.

2 Input data used in PVAAS

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

2.1 Determining suitability of assessments

2.1.1 Current assessments

In order to be used appropriately in any value-added analyses, the scales of these tests must meet three criteria. (Additional details on each of these requirements are provided in Section 8, Data quality and pre-analytic data processing, on page 34.)

- There is sufficient stretch in the scales to ensure that progress can be measured for both low-achieving students as well as high-achieving students. A floor or ceiling in the scales could disadvantage educators serving either low-achieving or high-achieving students.
- The test is highly related to the academic standards so that it is possible to measure progress with the assessment in that subject/grade/year.
- The scales are sufficiently reliable from one year to the next. This criterion typically is met when there are a sufficient number of items per subject/grade/year, and this will be monitored each subsequent year that the test is given.

These criteria are met by Pennsylvania's standardized assessments.

The current value-added implementation includes assessments measuring Pennsylvania's standards (PSSA and Keystones). There is potential to provide value-added reporting based on college and career readiness assessments.

2.1.2 Transitioning to future assessments

Pennsylvania transitioned to new assessments based on new standards in the 2014-2015 school year. Changes in testing regimes occur at regular intervals within any state, and these changes need not disrupt the continuity and use of value-added reporting by educators and policymakers. Based on twenty years of experience with providing value-added and growth reporting to educators, SAS has developed several ways to accommodate changes in testing regimes.

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

2.2 Assessment data used in Pennsylvania

The state tests are administered in the spring semester except for the Keystone assessments, which are given in the summer, winter, and spring semesters.

1.1.1 Tests given in consecutive grades for the same subject

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

- Pennsylvania System of School Assessment (PSSA) mathematics in grades three through eight.
- PSSA English Language Arts in grades three through eight.

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

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

- PSSA science in grades four and eight.
- Pennsylvania Keystone assessments in Algebra I, Biology, and Literature.

2.2.2 Student identification information

The following information is received by EVAAS from PDE:

- Student last name
- Student first name
- Middle initial (if available)
- Student date of birth
- PA secure ID

2.2.3 Assessment information provided

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

- Scale score
- Performance level
- Test taken
- Tested grade
- Tested semester
- District AUN
- School code

2.2.4 General exclusion of non-public school assessment data

No student scores are utilized from any non-public schools. These include prior year scores for students that are now in public schools.

2.3 Student-level information

Student-level information is used in creating the web application to assist educators analyze the data to inform practice and assist all students with academic progress. EVAAS receives this information in the

form of various socioeconomic, demographic, and programmatic identifiers in the student data system. Currently, these categories are as follows:

- Gifted IEP – added with enrollment data
- Service Plan 504 – added with enrollment data
- Migrant
- Limited English Proficient
- Economically Disadvantaged
- Special Ed IEP
- Gender
- Historically Underperforming
- Full Academic Year
- Foreign Exchange
- English Language Learner – first year
- Race
 - American Indian/Alaskan Native
 - Asian/Pacific Islander (prior to 2011)
 - Asian (beginning in 2011)
 - Native Hawaiian or Other Pacific Islander (beginning in 2011)
 - Black, Non-Hispanic
 - Hispanic
 - White, Non-Hispanic
 - Multi-Racial

2.4 Teacher-level information

A high level of reliability and accuracy is critical for using value-added measures for both improvement purposes and high stakes decision-making. Before teacher-level value-added measures are calculated, teachers in Pennsylvania are given the opportunity to complete roster verification to verify *linkages* between themselves and their students during the year. Roster verification by the individual teachers is an important part of a valid system. Roster verification enables teachers to confirm their class rosters for students they taught for a particular subject, grade, and year. These linkages, or records of teacher responsibility for specific students in specific subjects and grades, are verified by administrators at the school and district levels as an additional check. The roster verification process also captures different teaching scenarios where multiple teachers can share instruction. Verification therefore makes teacher-level analyses much more reliable and accurate.

EVAAS provides a roster verification application embedded within PVAAS. To initialize the linkages that will be verified, EVAAS receives data from the Pennsylvania Information Management System (PIMS) Staff Student Subtest template provided by PDE that contains a record for each teacher/student instructional relationship for each assessment. The roster verification process refines this data and the percentage of instructional responsibility of each teacher that may be attributed to a student by allowing teachers and administrators to go in and modify and verify this information.

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

- District AUN
- District name
- School code
- School name
- Teacher level identification
 - Teacher name
 - PPID
- Student linking information, including PAsecureID
- Subjects
- Semester
- Percentage of Student + Teacher enrollment
- Percentage of Full/Partial Instruction

As teachers across the Commonwealth participate in the roster verification process, the last two pieces of information are modified as needed. The first is the percentage of student + teacher enrollment, which is the percentage that a teacher and student are currently enrolled with one another – from day one of the subject/grade/course through the last instructional day before the testing window for that subject/grade/content area. The second is the percentage of full/partial instruction, which captures information regarding team teaching or shared instruction between two or more teachers. These two percentages are determined by LEAs. Both of these pieces of information are multiplied together to obtain an overall percentage of instructional responsibility.

3 Value-added analyses

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

- Growth = current achievement/current results compared to all prior achievement/prior results, with achievement being measured by a quality assessment such as the PSSA and Keystone exams.

In practice, growth must be measured using an approach that is sophisticated enough to accommodate many non-trivial issues associated with student testing data. Such issues include students with missing test scores, students with different entering achievement, and measurement error in the test. In Pennsylvania, EVAAS provides two main categories of value-added models, each comprised of district-, school-, and teacher-level reports.

- **Multivariate Response Model (MRM)** is used for tests given in consecutive grades, like the PSSA math and English Language Arts assessments in grades three through eight.
- **Univariate Response Model (URM)** is used when a test is given in non-consecutive grades, such as PSSA science assessments in grades four and eight or any Keystone exams.

Both models offer the following advantages:

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

Each model is described in greater detail in the following sections, in both conceptual terms and the standard technical notation used by statisticians. The two types of models are specific implementations of common statistical models, which have been used for decades in a number of other industries such as pharmaceutical and medical research. Simply put, these general models are well suited for identifying relationships in large, complex data sets, such is the case with Pennsylvania student testing records.

In these models, as a result of using all available test scores and including students, even if they have missing test scores, it is not necessary to make *direct* adjustments for students' background characteristics. In short, these adjustments are not necessary because each student serves as his or her own control. To the extent that socioeconomic/demographic influences persist over time, these influences are already represented in the student's data. As a 2004 study by The Education Trust stated, specifically with regards to the SAS EVAAS modeling:

“[I]f a student's family background, aptitude, motivation, or any other possible factor has resulted in low achievement and minimal learning growth in the past, all that is taken into account when the system calculates the teacher's contribution to student growth in the present.”

Source: Carey, K (Winter 2004). *The Real Value of Teachers: If Teachers Matter, Why Don't We Act Like It?* (The Education Trust: Washington DC).

In other words, while technically feasible, adjusting for student characteristics in sophisticated modeling approaches is not necessary from a statistical perspective; and the value-added reporting in Pennsylvania does not make any direct adjustments for students' socioeconomic/demographic

characteristics. Through this approach, Pennsylvania avoids the problem of building a system that creates differential expectations for groups of students based on their backgrounds.

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

3.1 Multivariate Response Model (MRM) reporting for tests in consecutive grades

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

The MRM is a *gain-based model*, which means that it measures growth between two points in time for a group of students. The growth expectation is met when a cohort of students from grade to grade maintains the same relative position with respect to statewide student achievement in that year for a specific subject and grade.

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

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

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

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

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

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

3.1.1 MRM at the conceptual level

An example data set with some description of possible value-added approaches may be helpful for conceptualizing how the MRM works and why a simple approach to measuring growth is problematic with missing test scores from students. Assume that ten students are given a test in two different years with the results shown in [Table 1](#). The goal is to measure academic growth (gain) from one year to the next. Two simple approaches are to calculate the mean of the differences *or* to calculate the differences of the means. When there is no missing data, these two simple methods provide the same answer (5.80 on the left in [Table 1](#)); however, when there is missing data, each method provides a different result (9.57 vs. 3.97 on the right in [Table 2](#)). A more sophisticated model is needed to address this problem.

Table 1: Scores without missing data

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

Table 2: Scores with missing data

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

The MRM uses the correlation between current and previous scores in the non-missing data to estimate a mean for the set of all previous and all current scores as if there were no missing data. It does this *without* explicitly imputing values for the missing scores. This means that the model does not enter

estimated test scores for students who are missing test scores. In other words, the model does not make assumptions about students' missing test scores. The model can avoid imputation by measuring gains, or the differences between estimated means.

In the tables, the difference between these two estimated means is an estimate of the average gain for this group of students. In this small example, the estimated difference is 5.8. Even in a small example such as this, the estimated difference is much closer to the difference with no missing data than either measure obtained by the mean of the differences (9.57) or difference of the means (3.97). This method of estimation has been shown, on average, to outperform both of the simple methods.¹ In this small example, there were only two grades and one subject. Larger data sets, such as those used in actual EVAAS analyses for Pennsylvania, provide better correlation estimates by having more student data, subjects, and grades, which in turn provide better estimates of means and gains.

This small example is meant to illustrate the need for a model that will accommodate incomplete data and provide a reliable measure of progress. It represents the conceptual idea of what is done with the school and district models. The teacher model is slightly more complex, and all models are explained in more detail below (in Section 3.1.3). The first step in the MRM is to define the scores that will be used in the model.

3.1.2 Normal curve equivalents

3.1.2.1 Why EVAAS uses normal curve equivalents in MRM

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

NCEs are on a familiar scale because they are scaled to look like percentiles. However, NCEs have a critical advantage for measuring growth: they are on an equal-interval scale. This means that for NCEs, unlike percentile ranks, the distance between 50 and 60 is the same as the distance between 80 and 90. NCEs are constructed to be equivalent to percentile ranks at 1, 50, and 99, with the mean being 50 and the standard deviation being 21.063 by definition. Although percentile ranks are usually truncated above 99 and below 1, NCEs are allowed to range above 100 and below 0 to preserve their equal-interval property and to avoid truncating the test scale. For example, in a typical year in Pennsylvania, the average maximum NCE is approximately 118. For display purposes in the PVAAS web application, NCEs are shown as integers from 1-99. Truncating would create an artificial ceiling or floor which may bias the results of the value-added measure for certain types of students forcing the gain to be close to 0 or even negative.

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

¹ See, for example: Wright, S. P. (2004), “Advantages of a Multivariate Longitudinal Approach to Educational Value- Added Assessment without Imputation,” Paper presented at National Evaluation Institute, online at <https://pvaas.sas.com/support/EVAAS-AdvantagesOfAMultivariateLongitudinalApproach.pdf>

By definition, the mean (or average) NCE score for the reference distribution is 50 for each grade and subject. “Growth” is the difference in NCEs from one year/grade to the next in the same subject. The growth standard, which represents a “normal” year’s growth, is defined by a value of zero. More specifically, it maintains the same position in the reference distribution from one year/grade to the next. **It is important to reiterate that a gain of zero on the NCE scale does not indicate “no growth.” Rather, it indicates that a group of students in a district, school, or classroom has maintained the same position in the state distribution from one grade to the next.** The expectation of growth can be set differently by using a reference distribution to create NCEs or by using each individual year to create NCEs. For more on Growth Expectation, see Section 4 on page 23.

1.1.1.3 How EVAAS uses normal curve equivalents in MRM

There are multiple ways of creating NCEs. EVAAS uses a method that does *not* assume the underlying scale is normal since experience has shown that some testing scales are not normally distributed and this will ensure an equal interval scale. Table 3 provides an example of the way that EVAAS converts scale scores to NCEs.

The first five columns of Table 3 show an example of a tabulated distribution of test scores from Pennsylvania data. The tabulation shows, for each possible test score, in a particular subject, grade, and year, how many students made that score (“Frequency”) and what percent (“Percent”) that frequency was out of the entire student population (in Table 3 the total number of students is approximately 130,000). Also tabulated are the cumulative frequency (“Cum Freq,” which is the number of students who made that score or lower) and its associated percentage (“Cum Pct”).

The next step is to convert each score to a percentile rank, listed as “Ptile Rank” on the right side of Table 3. If a particular score has a percentile rank of 48, this is interpreted to mean that 48% of students in the population had a lower score and 52% had a higher score. In practice, a non-zero percentage of students will receive each specific score. For example, 2.2% of students received a score of 1368 in Table 3. The usual convention is to consider half of that 2.2% to be “below” and half “above.” Adding 1.1% (half of 2.2%) to the 39.9% who scored below the score of 1368 produces the percentile rank of 41.0 in Table 3.

Table 3: Converting tabulated test scores to NCE values

Score	Frequency	Cum Freq	Percent	Cum Pct	Ptile Rank	Z	NCE
1340	2,820	48,620	2.2	37.6	36.6	-0.344	42.76
1354	2,942	51,562	2.3	39.9	38.8	-0.285	44.00
1368	2,880	54,442	2.2	42.2	41.0	-0.226	45.23
1382	2,954	57,396	2.3	44.4	43.3	-0.169	46.45
1396	3,064	60,460	2.4	46.8	45.6	-0.110	47.69
1411	2,982	63,442	2.3	49.1	48.0	-0.051	48.93
1425	3,166	66,608	2.5	51.6	50.4	0.009	50.19

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

3.1.3 Standard statistical notation of the linear mixed model and the MRM

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

of that linear function. The standard error of the estimable linear function can be found by taking the square root of this quantity.

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

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

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

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

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

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

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

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

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

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

3.1.3.1 District and school level

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

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

the subjects and grades for which the student has test scores. In this way, the MRM is able to use all available scores from each student.

Algebraically, the district MRM is represented as:

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

where y_{ijkl} 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_{ijkl} is the estimated mean score for this particular district, subject, grade, and year. ϵ_{ijkl} is the random deviation of the i^{th} student's score from the district mean.

The school MRM is represented as:

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

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

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

Solving the mixed model equations for the district or school MRM produces a vector b that contains the estimated mean score for each school (in the school model), subject, grade, and year. To obtain a value-added measure of average student growth, a series of computations can be done using the students from a school in a particular year and all of their prior year schools. Because students may change schools from one year to the next (in particular when transitioning from elementary to middle school, for example), the estimated mean score for the prior year/grade utilizes a weighted average of schools that fed students into the school, grade, subject, and year in question. Prior year schools are not utilized if they are feeding students in very small amounts (less than 5) since those students likely do not represent the overall achievement of the school that they are coming from. For certain schools with very large rates of mobility, the estimated mean for the prior year/grade only includes students that existed in the current year. Mobility is taken into account within the model so that growth of students is computed using all students in each school, including those that may have moved buildings from one year to the next.

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

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

3.1.3.2 Teacher-level

The teacher estimates use a more conservative statistical process to lessen the likelihood of misclassifying teachers. Each teacher is assumed to be the state average in a specific year, subject, and

grade until the weight of evidence pulls him or her either above or below that state average. Furthermore, the teacher model is a “layered” model, which means that:

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

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

To allow for the possibility of many teachers with relatively few students per teacher, MRM enters teachers as random effects via the Z matrix in the linear mixed model. The X matrix contains a column for each subject/grade/year, and the b vector contains an estimated state mean score for each subject/grade/year. The Z matrix contains a column for each subject/grade/year/teacher, and the u vector contains an estimated teacher effect for each subject/grade/year/teacher. The R matrix is as described above for the district or school model. The G matrix contains teacher variance components, with a separate unique variance component for each subject/grade/year. To allow for the possibility that a teacher may not receive similar growth measures in different subjects and grades, 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 σ^2_{jkl} is the teacher variance component for the j^{th} subject in the k^{th} grade in the l^{th} year, and I is an identity matrix.

Algebraically, the teacher model is represented as:

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

y_{ijkl} is the test score for the i^{th} student in the j^{th} subject in the k^{th} grade in the l^{th} year. $\tau_{ijk^*l^*t}$ is the teacher effect of the t^{th} teacher on the i^{th} student in the j^{th} subject in grade k^* in year l^* . The complexity of the parenthesized term containing the teacher effects is due to two factors. First, in any given subject/grade/year, a student may have more than one teacher. The inner (rightmost) summation is over all the teachers of the i^{th} student in a particular subject/grade/year. $\tau_{ijk^*l^*t}$ is the effect of those teachers. $w_{ijk^*l^*t}$ is the fraction of the i^{th} student’s instructional time claimed by the t^{th} teacher. Second, as mentioned above, this model allows teacher effects to accumulate over time. That is, how well a student does in the current subject/grade/year depends not only on the current teacher but also on the accumulated knowledge and skills acquired under previous teachers. The model will take into account the patterns that persist as students move on to different classrooms and will refine estimates of growth in the current year. 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 district and school estimates, the value-added estimates for teachers are not calculated by taking differences between estimated mean scores to obtain mean gains. Rather, this teacher model produces teacher “effects” (in the u vector of the linear mixed model). It also produces, in the fixed-effects vector b , state-level mean scores (for each year, subject and grade). Because of the way the X and Z matrices are encoded, in particular because of the “layering” in Z , teacher gains can be

estimated by adding the teacher effect to the state mean gain. That is, the interpretation of a teacher effect in this teacher model is as a gain, expressed as a deviation from the average gain for the state in a given year, subject, and grade.

On the following page, Table 4 illustrates how the Z matrix is encoded for three students who have three different scenarios of teachers during grades three, four, and five in two subjects, math (M) and ELA (E). Teachers are identified by the letters A–F.

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

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

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

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

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

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

Table 4: Encoding the Z matrix

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

3.1.4 Where the MRM is used in Pennsylvania

The MRM is used with the PSSA test in math and English Language Arts in grades three through eight. All of this data is used in each of the three separate analyses to obtain value-added measures at the district, school, and teacher level in grades four through eight.

The MRM methodology provides estimated measures of progress for up to three years in each subject/grade/year for district, school and teacher analyses provided that the minimum student requirements are met. For each subject, measures are also given across grades (4-8), across years (three year averages), as well as combined across years and grades.

At the teacher level, value-added measures for each PSSA subject/grade (4-8)/year are computed (and displayed on the PVAAS password protected web application available at <https://pvaas.sas.com/>).

More information regarding teacher level composite measures that use all teacher level data from up to three consecutive years can be found in Section 6 on page 30.

3.1.5 Students included in the analysis

In general, all of every student's math and ELA PSSA results are incorporated into the models. Some student scores may be excluded if they are flagged as outliers or due to the other business rules described in Section 8. In addition, exclusion rules are described below for the different levels of the analysis. Note that, in the MRM, students assessed in the most recent school year are included even if

they do not have prior testing history, and the conceptual explanation provided in Section 3.1.1 outlines why this provides a more reliable estimate of growth than excluding students simply for having missing test scores.

3.1.5.1 District and school level

The analyses for schools and districts include all students testing on the PSSA math and/or ELA tests. Students that are not enrolled for a full academic year (FAY) are not included in the analysis. Student scores that may be considered outliers are not used in the analysis. Students that are considered ELL-first year or foreign exchange are also not included in the analysis.

3.1.5.2 Teacher-level

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

Students are excluded from the teacher analysis if the students are not claimed by the teacher for at least 10% of their instructional responsibility. Because of this, the FAY designation is not used to exclude students from the analysis. Students are still excluded at this level if they are considered ELL-first year or foreign exchange. Student scores that may be considered outliers are not used in the analysis.

3.1.6 Minimum number of students for reporting

3.1.6.1 District and school level

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

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

- There must be at least eleven students who are associated with the school or district in that subject/grade/year.
- There is at least one student at the school or district who has a “simple gain,” which is based a valid test score in the current year/grade as well as the prior year/grade in the same subject.
- Of those students who are associated with the school or district in the current year/grade, there must be at least five students that have come from any other school for that prior school to be used in the gain calculation. This ensures that the prior school is representative of the students included in the model.

3.1.6.2 Teacher-level

The teacher-level value-added model includes teachers who are linked to at least eleven students with a valid test score in the same subject and grade. This requirement does not consider the percentage of instructional responsibility that the teacher has with each student in a specific subject/grade.

However, in order to receive a teacher value-added report for a particular year, subject and grade, a teacher must have at least six Full Time Equivalent (FTE) students in a specific subject/grade/year. The teacher’s number of FTE students is based on the number of students linked to that teacher and the overall percentage of instructional responsibility the teacher has for each student. For instance, if a teacher taught ten students and claimed 50% of their instructional responsibility, then the teacher’s FTE

number of students would be five and the teacher would not receive a teacher value-added report. If another teacher taught twelve students and claimed 50% of their instructional responsibility, then that teacher would have six FTE students and that teacher would receive a teacher value-added report. The instructional responsibility attribution is obtained from the linkage roster verification process that is in use in PVAAS. This information is outlined in Section 2.

Students are linked to a teacher based on the subject/grade/content area taught and the state assessment taken. In some cases, the course being taught may not directly align to all state assessments taken by the student and, in those cases, linkage by EVAAS is not mandatory in accordance with PDE policy. For example, all eighth grade students take the PSSA mathematics grade 8 test. However, some eighth grade students (as well as students in younger grades) are actually also enrolled in a Keystone Algebra I course in a PSSA-tested grade rather than the general grade 8 math course. Their teachers will not be automatically linked in the PVAAS roster verification system to these eighth grade students enrolled in Algebra I by EVAAS unless the LEA submits these links to the PSSA Math assessment into the PIMS Staff-Student-Subtest template. As a result, these teachers may not receive a PSSA Math Grade 8 report. LEAs are responsible to make this determination. LEAs *make the choice as to whether* to have their teachers linked to such students if the students should be included in the teacher's PSSA Math Grade 8 value-added report. If the LEA determines that the Algebra I teacher also has responsibility for eighth grade students on the grade 8 assessment (or grade 7, etc.), then that teacher would receive a grade eight mathematics report based on the students who took the PSSA Math Assessment *and* a separate Algebra I report through both state assessments.

The process for creating an accurate link between students and teachers (roster verification) allows teachers and principals to review the attribution used in the PVAAS reports. For more information about teacher roster verification, email pdepvaas@iu13.org.

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

Tests that are not given for consecutive years require a different modeling approach from the MRM, and this modeling approach is called the univariate response model (URM). The statistical model can also be classified as a linear mixed model and can be further described as an analysis of covariance (ANCOVA) model. The URM is a regression-based model, which measures the difference between students' predicted scores for a particular subject/year with their observed scores. The growth expectation is met when students with a district/school/teacher made the same amount of progress as students in the average district/school/teacher with the state for that same year/subject/grade.

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

- It does not require students to have all predictors or the same set of predictors, so long as a student has at least three prior test scores in any subject/grade.
- It minimizes the influence of measurement error by using all prior data for an individual student. Analyzing all subjects simultaneously increases the precision of the estimates.
- It allows educators to benefit from all tests, even when tests are on differing scales.
- It accommodates teaching scenarios where more than one teacher has responsibility for a student's learning in a specific subject/grade/year.

In Pennsylvania, URM value-added reporting is available for the PSSA science test in grades four and eight at the district, school, and teacher levels. This reporting is also available for the Keystone exams in Algebra I, Biology, and Literature at the district, school, and teacher levels.

3.2.1 URM at the conceptual level

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

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

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

3.2.2 Standard statistical notation of the district, school and teacher models

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.2.3 Students included in the analysis

In order for a student's score to be used in the district, school, or teacher level analysis for a particular subject/grade/year, the student must have at least three valid predictor scores that can be used in the analysis, all of which cannot be deemed outliers. The only exception is in grade 4 science, where students need at least two valid predictors. These scores can be from any year, subject, and grade that are used in the analysis. It will include subjects other than the subject being predicted. The required three predictor scores are needed to sufficiently dampen the error of measurement in the tests to provide a reliable measure. If a student does not meet the three score minimum, then that student is excluded from the analyses. 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. For Keystone assessments, only students' "admin scale scores" in the DRC Keystone student level file are considered for inclusion in the analyses.

A student is only included in the district or school level analysis if they have met the full academic year requirement and they are not considered ELL-first year or foreign exchange.

For the teacher-level analysis, students must have been claimed by a teacher at least 10% of their instructional responsibility to be included. Students must not be considered ELL-first year or foreign exchange if included in the teacher level analysis.

The reporting year includes the summer administration, winter administration, and the spring administration of each reporting year. Summer administration is considered to be the first administration of the school year with the reporting results provided in the fall, followed by the winter and spring administrations.

3.2.3.1 Special Considerations for Keystone assessments in the district and school level analysis

Starting with SY2015-16 reporting, the following business rules will be used to determine which student test scores are included in the Keystone models. They will be applied in the listed order.

1. Any scores for a student that are reported after the first time a student is proficient or advanced in that subject are removed from consideration for inclusion in the model. As per PDE policy students are not to be retesting on a Keystone exam if they are already proficient or higher.
2. Any scores for a student that are lower than a previously reported score (in the current year or any prior years) in that Keystone subject for that student are removed from consideration for inclusion in the analyses.
3. Of the remaining scores, only the highest score within *that* reporting year is kept for *possible* inclusion in the analyses. If the student's highest score is *not* during the current school/reporting year, the student would not be included in the current year of value-added analyses.
4. After steps one through three are applied, the remaining student scores may be removed from the model for the following reasons:

- a. Students that did not meet the full academic year requirement are removed.
- b. Students in their first year of ELL are removed.
- c. Foreign exchange students are removed.
- d. Outliers are removed.
- e. Students who have an IEP and test outside of their district of residence.

3.2.4 Minimum number of students for reporting

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

Also according to PDE policy, for teacher level reporting, there must be eleven student scores in that year, subject and grade that have the required three prior test scores needed to obtain a predicted score in that year, subject and grade. Again, in order to receive a teacher value-added report for a particular year, subject and grade, a teacher must have at least six Full Time Equivalent (FTE) students in a specific subject/grade/year as described in Section 3.1.6.2.

3.2.5 Prior tests used in the URM analyses

As mentioned above, the URM is run for each individual year, subject, and grade (where appropriate). When examining the relationship with prior test data, only certain tests are included. These are the following:

- Prior PSSA math, ELA, and science are used as prior testing history for PSSA science
- Prior PSSA math, ELA, and science are used as prior testing history for Keystone Algebra I
- Prior PSSA math, ELA, science, and Keystone Algebra I when available are used as prior testing history for Keystone Biology
- Prior PSSA math, ELA, science, Keystone Algebra I when available, and Keystone Biology when available are used as prior testing history for Keystone English Literature

Note that prior Keystone scores are not used as predictors in the same Keystone subject. For example, Algebra I is not used as a predictor for Algebra I repeaters. With the URM, test scores are only used as predictors if at least half of the students with the current year test scores have that as a prior measure. To date, there have not been enough students who take the test for a second time to use the scores as predictors, as it is generally a much smaller subset of students who take a Keystone and have taken that prior test in prior years. Since this is a statewide model that uses all students across the state and all of their prior testing to set a predicted score for each student, these prior scores are not considered in the model.

4 Growth expectation

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

- Growth = current achievement/current results compared to all prior achievement/prior results; with achievement being measured by a quality assessment such as the PSSA and Keystone tests.

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

However, the definition of “expected growth” varies by model, and the precise definition depends on the selected model and state preference, and this section provides more details on the options and selections for defining expected growth. This document describes the expected growth as either a “base year” or “intra-year” approach. Base year refers to a growth expectation that is based on a particular year, say 2006, and any growth in the current year will be compared to the distribution of student scores in the base year. Currently, Pennsylvania uses an intra-year approach because of testing transitions. Intra-year refers to a growth expectation that is always based on the current year (2012 for 2012 growth estimates, 2013 for 2013 growth estimates, and so on).

In years prior to 2013, the base year approach was used with 2006 as the base year. Because this was used in Pennsylvania for many years, this approach will also be fully described. The change to the intra-year approach was done to accommodate the change to PA core standards.

4.1 Intra-year approach

4.1.1 Description

This approach is used currently in Pennsylvania for all value-added measures and must be used in the MRM reporting during the transition to new assessments and the concept is always used in the URM reporting. The actual definitions in each model are slightly different, but the concept can be considered as the average amount of progress seen across the state in a statewide implementation.

Using the URM model the definition of the expectation is that students with a district, school, or teacher made the same amount of progress as students with the average district, school, or teacher in the state for that same year/subject/grade. If not all students are taking an assessment in the state, then it may be a subset.

Using the MRM model, the definition of this type of expectation of growth is that students maintained the same relative position with respect to the statewide student achievement from one year to the next in the same subject area. As an example, if students’ achievement was at the 50th NCE in 2014 grade four math, based on the 2014 grade four math statewide distribution of student achievement, and their achievement is at the 50th NCE in 2015 grade five math, based on the 2015 grade five math statewide distribution of student achievement, then their estimated gain is 0.0 NCEs.

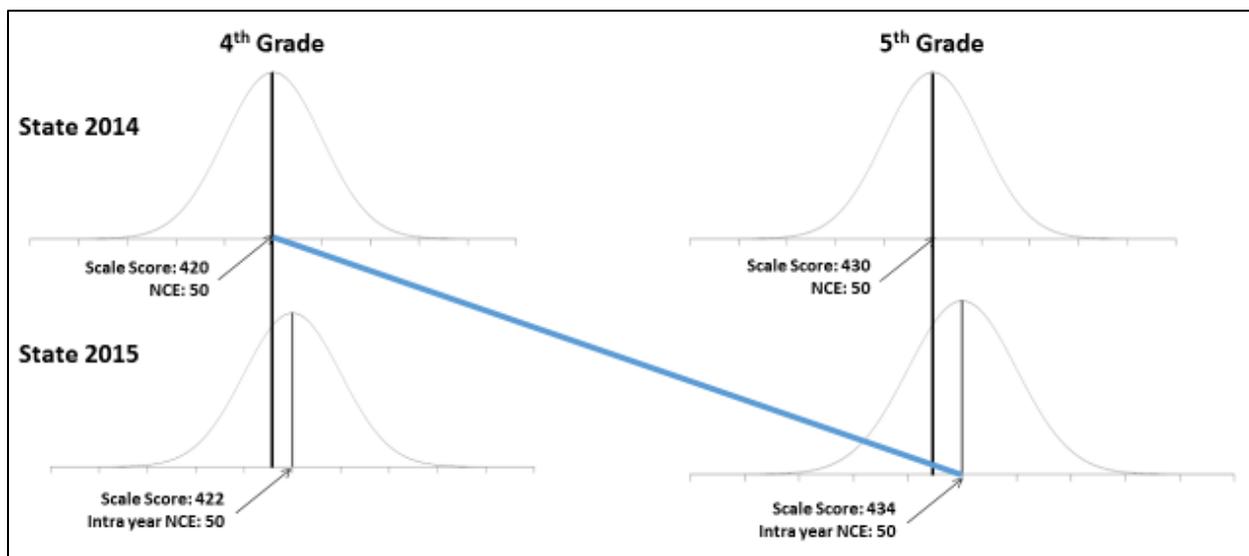
With this approach, the value-added measures tend to be centered on the growth expectation every year, with approximately half of the district/school/teacher estimates above zero and approximately half of the district/school/teacher estimates below zero. This does not mean half would be in the positive and negative categories since many value-added measures are indistinguishable from the expectation when considering the statistical certainty around that measure.

4.1.2 Illustrated example

The graphic below (Graph 1) provides a *simplified* example of how growth is calculated with an intra-year approach when the state or pool achievement increases. The graphic below has four graphs, each of which plot the NCE distribution of scale scores for a given year and grade. In this example, the first year is 2014, and the graphic shows how the gain is calculated for a group of 2014 grade four students as they become 2015 grade five students. In 2014, our grade four students score, on average, 420 scale score points on the test, which corresponds to the 50th NCE (similar to the 50th percentile). In 2015, the students score, on average, 434 scale score points on the test, which corresponds to a 50th NCE *based on the 2015 grade five distribution of scores*. The 2015 grade five distribution of scale scores was higher than the 2014 grade five distribution of scale scores, which is why the lower right-hand graph is shifted slightly to the right. The blue line shows what is required for students to make expected growth, which would be to maintain their position at the 50th NCE in 2014 grade four as they become 2015 grade five students. The growth measure for these students is 2015 NCE – 2014 NCE, which would be 50 – 50 = 0. Similarly, if a group of students started at the 35th NCE, the expectation is that they would maintain that 35th NCE.

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

Graph 1: Intra-year approach example



4.2 Base year approach

4.2.1 Description

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

As a simplified example, if students' achievement was at the 50th NCE in 2006 grade four math, based on the 2006 grade four math scale score distribution, and at the 52nd NCE in 2007 grade five, based on the 2006 grade five math scale score distribution, then their estimated mean gain is 2 NCEs.

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

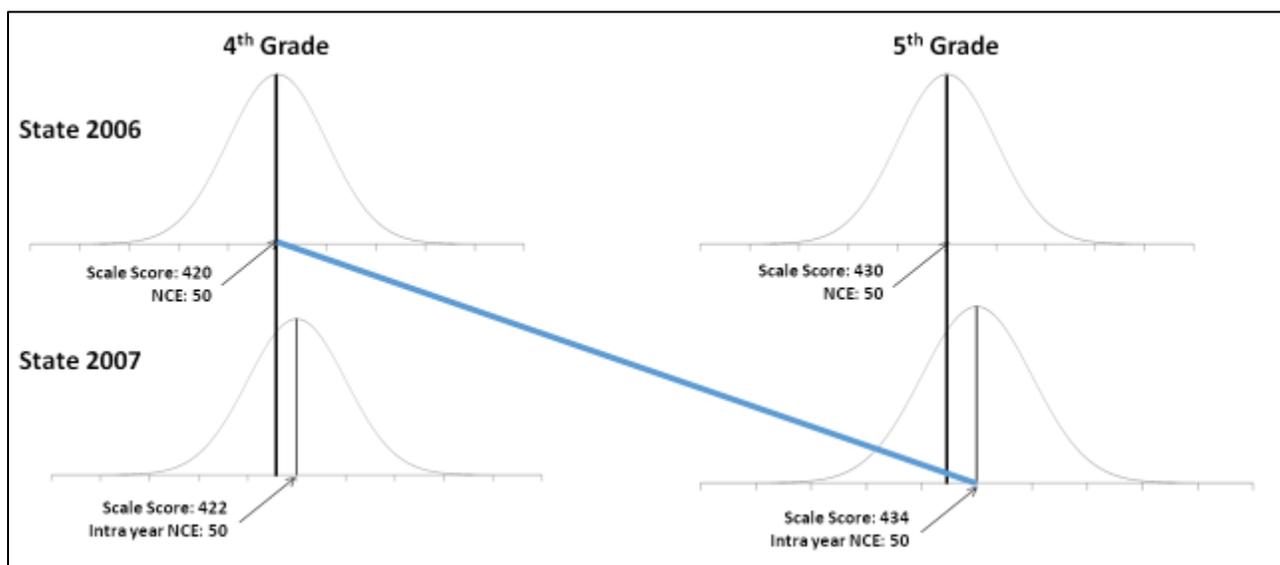
Following the implementation of any new assessments and changes in academic standards, the base year should be reset to an intra-year approach in order to accommodate the differences between the old and new testing regimes and minimize any impact on the value-added reporting. To be more specific, use of the intra-year approach is required if there is no mapping from the old assessment's scale to the new assessment's scale. However, even if that mapping does exist, the intra-year approach should be used to prevent any unusual swings in value-added measures.

4.2.2 Illustrated example

The graphic below (Graph 2) provides a *simplified* example of how growth is calculated with a base year approach when the state achievement increases. The graphic below has four graphs, each of which plot the NCE distribution of scale scores for a given year and grade. In prior years in Pennsylvania, the base year was 2006, and the graphic shows how the gain is calculated for a group of 2006 grade four students as they become 2007 grade five students. In 2006, our grade four students score, on average, 420 scale score points on the test, which corresponds to the 50th NCE (similar to the 50th percentile). In 2007, the students score, on average, 434 scale score points on the test, which corresponds to a 52nd NCE *based on the 2006 grade five distribution of scores*. The 2007 grade five distribution of scale scores was higher than the 2006 grade five distribution of scale scores, which is why the lower right-hand graph is shifted slightly to the right. The blue line shows what is required for students to make expected growth, which would be to maintain their position at the 50th NCE in 2006 grade four as they become 2007 grade five students. The growth measure for these students is 2007 NCE – 2006 NCE, which would be 52 – 50 = 2. Similarly, if a group of students started out at the 35th NCE in 2006 grade four and then moved their position to the 37th NCE in 2007 grade five, they would have a gain of two NCEs as well.

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

Graph 2: Base year approach example



4.3 Defining the expectation of growth during an assessment change

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

However, with the MRM methodology, a base year approach presents challenges since it requires the scales to stay consistent over time. That said, with the intra-year approach, the scales from one year to the next may be completely different from one another. This method converts any scale to a relative position and can be used through an assessment change.

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

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

5.1 Using standard errors derived from the models

As described in the modeling approaches section, each model provides an estimate of growth for a district, school, or teacher in a particular subject/grade/year as well as that estimate's standard error. The standard error is a measure of the quantity and quality of student level data included in the estimate, such as the number of students and the occurrence of missing data for those students. Because measurement error is inherent in any growth or value-added model, *the standard error is a critical part of the reporting*. Taken together, the estimate and standard error provide the educators and policymakers with critical information regarding the certainty that students in a district, school, or classroom are making decidedly more or less than the expected progress. Taking the standard error into account is particularly important for reducing the risk of misclassification (for example, indicating that the teacher's group of students did not meet the PA Standard for Academic Growth when the group of students really did) for high-stakes usage of value-added reporting.

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

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

5.2 Defining evidence of growth in terms of standard errors

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

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

- Dark Blue is an indication that the Growth Measure is more than 2 standard errors above the standard for PA Academic Growth (0). There is significant evidence of exceeding the standard for PA Academic Growth.
- Light Blue is an indication that the Growth Measure is at least 1 but less than 2 standard errors above the standard for PA Academic Growth (0). There is moderate evidence of exceeding the standard for PA Academic Growth.

- Green is an indication that the Growth Measure is less than 1 standard error above the standard for PA Academic Growth (0) and no more than 1 standard error below it (0). There is evidence of meeting the standard for PA Academic Growth.
- Yellow is an indication that the Growth Measure is more than 1 but no more than 2 standard errors below the standard for PA Academic Growth (0). There is moderate evidence of not meeting the standard for PA Academic Growth.
- Red is an indication that the Growth Measure is more than 2 standard errors below the standard for PA Academic Growth (0). There is significant evidence of not meeting the standard for PA Academic Growth.

The terminology might be slightly different depending on what analysis is being categorized. For instance, teacher-level reporting uses the same boundary definitions, but the language is different to indicate the teacher-level analysis. In the reporting, there is a need to display the values that are used to determine these categories. This value is typically referred to as the growth index and is simply the estimate or mean gain divided by its standard error. ***Since the expectation of growth is zero, this measures the certainty about the difference of a growth measure to zero.***

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

5.3 Rounding and truncating rules

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

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

5.4 Other scales used for reporting in Pennsylvania

In order to combine the PVAAS 3-year rolling Average Growth Index (AGI) with the other multiple measures of the evaluation system, it is necessary to convert the PVAAS 3-year rolling AGI to a 0 to 3 scale. The following table illustrates this conversion. Values between the values displayed in the table are scaled linearly.

Table 5: For Teachers - Crosswalk of all rating tools

PVAAS Growth Color Indicator	PVAAS 3-Year Rolling Average Growth Index (AGI)	Teacher Rating 0 to 3 Scale	100-Point Scale
Dark Blue	3.00 or greater	3.00	100
Dark Blue	2.00 to 2.99	2.50 to 2.99	90.00 to 99.99
Light Blue	1.00 to 1.99	2.00 to 2.49	80.00 to 89.99
Green	-1.00 to 0.99	1.50 to 1.99	70.00 to 79.99
Yellow	-2.00 to -1.01	0.50 to 1.49	60.00 to 69.99
Red	-3.00 to -2.01	0.41 to 0.49	50.00 to 59.99
Red	-3.01 or less	0.40	49.00

Table 6: For Schools - AGI conversion to 100-point scale

If the AGI is	The Scale Score is
3.00 or greater	100
Less than 3.00 but greater than or equal to 1.00	$10 \times (\text{AGI} + 7)$, truncated to a whole number
Less than 1.00 but greater than or equal to -1.00	$5 \times (\text{AGI} + 15)$, truncated to a whole number
Less than -1.00 but greater than or equal to -3.00	$10 \times (\text{AGI} + 8)$, truncated to a whole number
Less than -3.00	50

NOTE: When an Average Growth Index falls exactly on the boundary between two ranges, the scale score conversion formula for the higher range is assigned.

6 Teacher multi-year composite calculation

The section captures how the policy decisions by PDE are implemented in the calculation of the composite for up to three consecutive school years for teachers in the tested subjects and/or grades.

6.1 Overview of teacher level composites

The following text provides a specific example of a teacher's composite, the key policy decisions can be summarized as follows:

- A multi-year trend composite is calculated using all subjects and grades for up to three consecutive school years.
- The composite for teachers weights each subject/grade/year equally.
- This multi-year trend will be calculated each year, but is not used in the Pennsylvania teacher evaluation until it contains three consecutive school years of value-added data. (Note: This does not need to be in the same state assessed subject/grade/content area.)

The composite for teachers will include PSSA math, ELA, science and any Keystone assessments. The following examples will be used to show how the up to three year composite is calculated for a sample teacher.

Table 7: Example of available data for PSSA multi-year composite for a sample teacher across subjects

Year	Subject	Grade	Value-Added Measure	Standard Error	Index
2014	Science	8	15.20	7.00	2.17
2014	Math	7	3.50	1.50	2.33
2015	ELA	8	0.50	1.40	0.36
2015	Math	8	4.50	1.60	2.81
2016	ELA	8	-0.30	1.20	-0.25
2016	Math	8	3.80	1.50	2.53

6.2 Calculating the index

For the teacher in the above example, they have taught a mixture of subjects and grades from 2014 to 2016. All of these measures will be utilized in the overall up to three year composite calculation. As explained in earlier sections, the model produces a value-added measure and standard error for each year/subject/grade possible for a teacher. These two values are used to see if there is statistical evidence that the value-added measure is different from the expectation of growth, which is zero.

In the above example, the value-added measures for math and ELA are on the NCE scale, whereas the value-added measure is reported in the scale score units in science. An index is calculated for each of these measures by dividing the value-added measure by its standard error and is given in the final column.

The index is standardized (unit-less) or in terms of the standard errors away from zero. This makes it possible to combine across subjects and grades. This standardized statistic has a standard error of 1.

6.3 Combining the index values across subjects, grades and years

To calculate the overall composite that uses value-added information for up to three years, the first step is to average the index values. In the above example, this would look like the following using the numbers from the last column of Table 7:

$$Avg. Index = \frac{1}{6}(2.17 + 2.33 + 0.36 + 2.81 - 0.25 + 2.53) = 1.66 \quad (20)$$

Since each of the individual index values have a standard error of 1, there needs to be an additional correction to recalculate the overall average index to make it have a standard error of 1 or so that it is standardized like the original index values. This uses a standard statistical practice to ensure the final index has a standard error of 1. This correction is simple, but to derive where it comes from, the standard error of an average index can be found using the following formula.

$$SE Avg. Index = \frac{1}{6}\sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = \frac{\sqrt{6}}{6} = \frac{1}{\sqrt{6}} \quad (21)$$

To calculate the new index, the average of the index values would be divided by the new standard error of the average index. Therefore, to get the new index value, the average of the indexes is multiplied by square root of the number of measures that went into it.

$$Composite Index = \frac{1.66}{\left(\frac{1}{\sqrt{6}}\right)} = 1.66 \sqrt{6} = 4.07 \quad (22)$$

7 PVAAS Projection Model

In addition to providing value-added modeling, PVAAS provides a variety of additional services including projected scores for individual students on tests the students have not yet taken or are not yet proficient (Keystones). These tests may include state-mandated tests (end-of-grade tests and end-of-course tests where available) as well as national tests such as college and career readiness exams (AP, PSAT, SAT and ACT). These projections can be used to predict a student's future success (trajectory to success) and so may be used to guide counseling and intervention to increase students' likelihood of future success. Table 8 below provides a list of which prior achievement scores are used to calculate specific projections.

Table 8: Prior achievement data used to calculate projection

Projection to...	Data used to calculate projection	Projected to/from
PSSA Math	PSSA Math and ELA	One to two grades above last tested grade
PSSA ELA	PSSA Math and ELA	One to two grades above last tested grade
PSSA Science	PSSA Math, ELA and Science (in grades available)	To the next science grade
Keystone Algebra I	PSSA Math, ELA and Science (in grades available)	Starting with those that last tested in grade five
Keystone English Literature and Biology	PSSA Math, ELA and Science (in grades available) and Algebra I when available	Starting with those that last tested in grade five
SAT, ACT, and AP	PSSA Math, ELA, Science, Keystones, and PSAT.	Starting with those that last tested in grade eight
PSAT	PSSA Math, ELA, Science, Keystones.	Starting with those that last tested in grade five

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

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \dots + \epsilon_i \quad (23)$$

The μ terms are means for the response and the predictor variables. α_j is the school effect for the j^{th} school, the school attended by the i^{th} student. The β terms are regression coefficients. Projections to

the future are made by using this equation with estimates for the unknown parameters (μ 's, β 's, sometimes α_j). The parameter estimates (denoted with "hats," e.g., $\hat{\mu}$, $\hat{\beta}$) are obtained using the most current data for which response values are available. The resulting projection equation for the i^{th} student is:

$$\hat{y}_i = \hat{\mu}_y \pm \hat{\alpha}_j + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots + \epsilon_i \quad (24)$$

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

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

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

$$Prob(\hat{y}_i \geq b) = \Phi\left(\frac{\hat{y}_i - b}{SE(\hat{y}_i)}\right) \quad (25)$$

8 Data quality and pre-analytic data processing

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

8.1 Data quality

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

8.2 Checks of scaled score distributions

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

8.2.1 Stretch

Stretch indicates whether the scaling of the test permits student growth to be measured for either very low- or very high-achieving students. A test “ceiling” or “floor” inhibits the ability to assess growth for students who would have otherwise scored higher or lower than the test allowed. There must be enough test scores at the high or low end of achievement for measurable differences to be observed. Stretch can be determined by the percentage of students who score near the minimum or the maximum level for each assessment. In 2015, the percentage of students who achieved a maximum score on the PSSA assessments was less than 0.05% across all subjects and grades. As an example, if a much larger percentage of students scored at the maximum in one grade compared to the prior grade, then it may seem that these students had negative growth at the very top of the scale. However, this is likely due to the artificial ceiling of the assessment. Percentages for all of the PSSA and Keystone assessments are well below acceptable values, meaning that the state tests have adequate stretch to measure value-added even in situations where the group of students are very high or low achieving.

8.2.2 Relevance

Relevance indicates whether the test is aligned with the curriculum. The requirement that tested material will correlate with standards if the assessments are designed to assess what students are expected to know and be able to do at each grade level. Since the Pennsylvania state assessments are designed to measure state curriculum, this is not an issue.

8.2.3 Reliability

Reliability can be viewed in a few different ways for assessments. Psychometrics view reliability as the idea that students would receive similar scores if they took the assessment multiple times. Reliability also refers to the assessment’s scales across years. Both of these types of reliability are important when measuring growth. The first type of reliability is important for most any use of standardized assessments. The second type of reliability is very important when a base year is used to set the

expectation of growth since this approach assumes that scale scores mean the same thing in a given subject and grade across years.

8.3 Data quality business rules

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

8.3.1 Missing grade levels

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

Of the 1763991 records from the 2015-2016 PSSA Math, ELA, and Science assessments, no records were excluded due to this business rule.

8.3.2 Duplicate (same) scores

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

Of the 2351518 records from the 2015-2016 PSSA Math, ELA, and Science and Keystone Algebra I, Biology and Literature assessments, 30 records (0.001%) were excluded due to this business rule.

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

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

Of the 2351518 records from the 2015-2016 PSSA Math, ELA, and Science and Keystone Algebra I, Biology and Literature assessments, no records were excluded due to this business rule.

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

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

Of the 2351518 records from the 2015-2016 PSSA Math, ELA, and Science and Keystone Algebra I, Biology and Literature assessments, 130 records (0.006%) were excluded due to this business rule.

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

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

Of the 1763991 records from 2015-2016 PSSA Math, ELA, and Science assessments, 9 records (less than 0.001%) were excluded due to this business rule.

8.3.6 Students with records that have unexpected grade level changes

If a student skips more than one grade level (e.g., moves from sixth in 2014 to ninth in 2015) or is moved back by one grade or more (i.e. moves from fourth in 2014 to third in 2015) 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.

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

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

8.3.8 Outliers

Student assessment scores are checked each year to determine if they are outliers in context with all of the other scores in a reference group of scores from the individual student. These reference scores are weighted differently depending on proximity in time to the score in question. Scores are checked for outliers using related subjects as the reference group. For example, when searching for outliers for math test scores, all math subjects (PSSA and Keystone) 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 PVAAS web application.

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

Considerations included in outlier detection are:

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

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

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

For low-end outliers, the rules are:

- The percentile of the score must be below 50.
- The t-value must be below -3.5 when looking at the difference between the score in question and the reference group of scores.
- The percentile of the comparison score must be above a certain value. This value depends on the position of the individual score in question but will range from 10 to 90 with the ranges of the individual percentile score.

For high-end outliers, the rules are:

- The percentile of the score must be above 50.
- The t-value must be above 4.0.
- The percentile of the comparison score must be below a certain value.
- There must be at least 3 scores in the comparison score average.

Of the 2351518 records from the 2015-2016 PSSA Math, ELA, and Science and Keystone Algebra I, Biology and Literature assessments, 781 records (0.03%) were excluded due to this business rule.