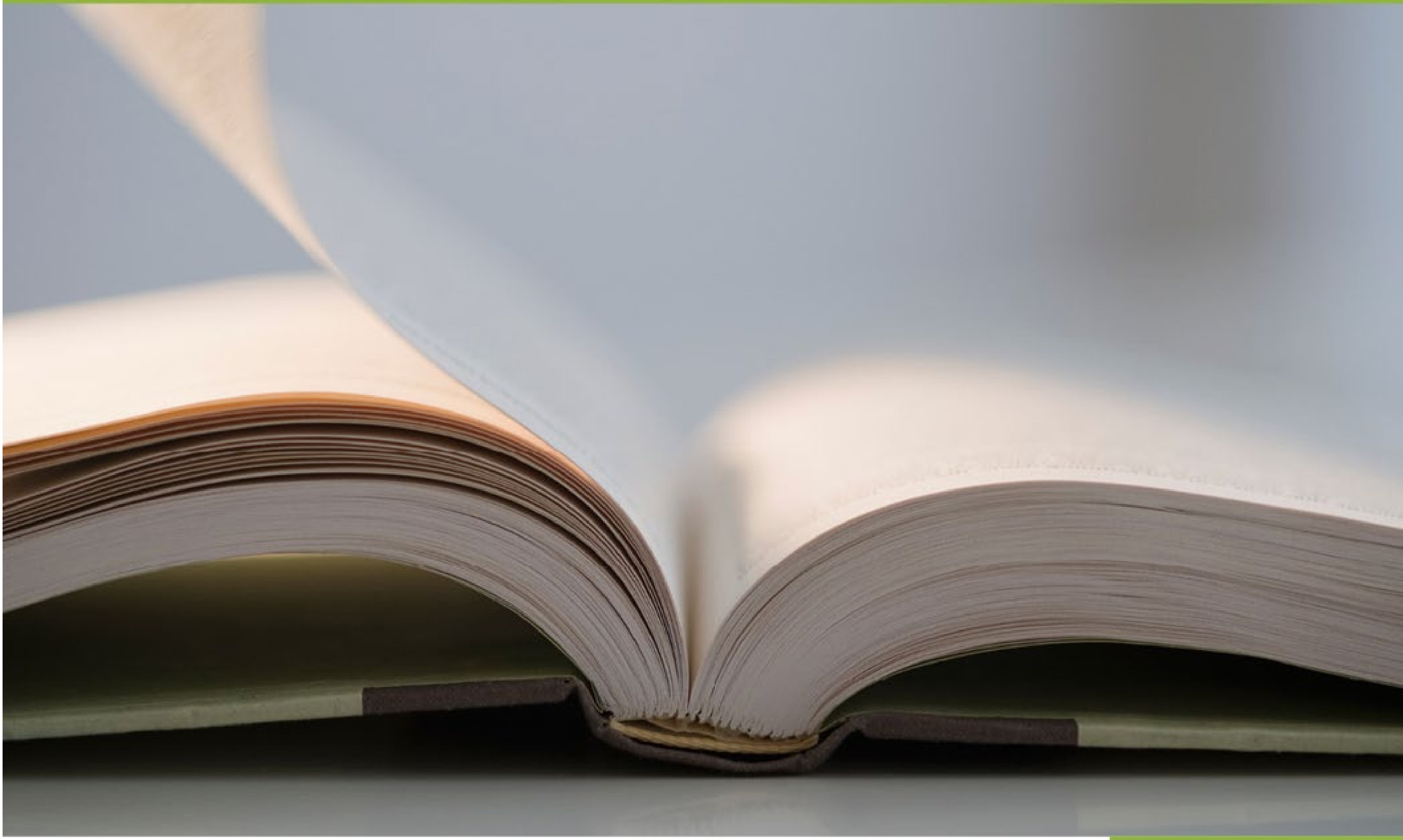
SAS® EVAAS

Statistical Models and Business Rules

Prepared for the Michigan Department of Education



Contents

1 Introduction to Growth Reporting in Michigan

In 2018, the Michigan Department of Education (MDE) made SAS EVAAS reporting available to its districts and schools. Teacher reporting was made available to any districts that wanted to opt in through the Michigan Data Hub (MiDataHub) project. Available through a secure web application, this reporting supports educators with school improvement with both reflective and proactive planning tools.

The core of EVAAS reporting is growth, which measures the change in achievement over time for a group of students. The change is based on student performance on a quality standardized assessment, such as M-STEP or MAP. EVAAS uses a set of growth (or value-added) models that have been available to districts, schools, and teachers in some states since 1993. When first implemented over two decades ago, EVAAS represented a paradigm shift for educators and policymakers to consider both achievement and growth rather than achievement alone. EVAAS reporting provides personalized feedback to districts, schools, and teachers and identifies the more (or less) effective practices in use. This insight can be leveraged to improve the academic experiences of their students.

Conceptually, growth is easy to understand; it is simply the change in achievement for a group of students over time. In practice, however, the implementation of a growth model is more complex. There are several decisions related to data, models, local policies and preferences, and business rules. The models themselves are highly sophisticated in order to address common concerns related to working with assessment data.

The purpose of this document is to guide you through the growth reporting based on the data, statistical models, policies, and practices selected by the Michigan Department of Education and currently implemented by SAS.

2 Data Inputs

This section provides details about the input data used in the Michigan growth models. In some cases, there might be additional student, teacher, school, and district information provided for various purposes.

2.1 Determining Suitability of Assessments

2.1.1 Current Assessments

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

- **There is sufficient stretch in the scales** to ensure that growth can be measured for both low-achieving students as well as high-achieving students. A floor or ceiling in the scales could disadvantage educators serving either low-achieving or high-achieving students.
- **The test is designed to assess the academic standards,** so it is possible to measure growth with the assessment in that subject, grade, and year. More information about Michigan academic standards can be found at the following link: https://www.michigan.gov/mde/0,4615,7-140-28753_64839_65510---,00.html
- **The scales are sufficiently reliable from one year to the next.** This criterion typically is met when there are a sufficient number of items per subject/grade/year, and this will be monitored each subsequent year that the test is given.

These criteria are monitored by SAS and MDE.

2.2 Assessment Data Used in Michigan

2.2.1 Assessments

For the 2018-19 school year, SAS received the following assessments for EVAAS reporting:

- M-STEP English Language Arts (ELA) and Mathematics in grades 3–7
- M-STEP Science in grades 5, 8, and 11, which are currently field tests
- M-STEP Social Studies in grade 5, 8, and 11
- PSAT 8/9 in Mathematics and ELA in grade 8
- PSAT 8/9 in Mathematics, Evidence-Based Reading and Writing in grade 9
- PSAT 10 in Mathematics, Evidence-Based Reading and Writing in grade 10
- SAT in Mathematics, Evidence-Based Reading and Writing in grade 11

These assessments are administered in the spring semester of the school year.

SAS received interim/benchmark assessments from districts that opted to submit them for EVAAS teacher reporting through MiDataHub, and the following assessments met the criteria for assessments in Section 2.1 as well as minimum number requirements in Section 3.1.6:

- MAP Mathematics in grades 1–8
- MAP Reading in grades 1–8

These assessments are administered at the beginning of year (BOY), middle of year (MOY) and end of year (EOY) for the 2017-18, 2018-19 and 2019-20 school years. BOY includes test scores from August through October, MOY includes test scores from December through February, and EOY includes test scores from March through June.

2.2.2 Student Identification and Assessment Information from MDE

SAS received the following student identification information from MDE:

- Student last name
- Student first name
- Student middle name
- Student date of birth
- Student state ID number (UIC)

SAS also received the following assessment information from MDE:

- Scale score
- Test taken
- Tested grade
- Tested subject
- Tested semester
- Tested performance level
- Full Academic Year designation
- Educational Entity Master District code
- Educational Entity Master District name
- Educational Entity Master School code
- Educational Entity Master School name

2.2.3 Student Identification and Assessment Information from MiDataHub for Interim/Benchmark Assessments

SAS used the following endpoints and required data elements from the MiDataHub for the interim/benchmark assessments:

- academicSubjectDescriptors
  - codeValue
  - description/shortDescription
  - namespace

- assessments
  - assessmentIdentifier
  - namespace
  - assessmentFamily
  - assessmentForm
  - assessmentTitle
  - assessmentVersion
  - academicSubjectDescriptor
  - gradeLevelDescriptor
- gradeLevelDescriptors
  - codeValue
  - description/shortDescription
  - namespace
- schools
  - schoolId
  - nameOfInstitution
  - operationalStatusDescriptor
  - schoolTypeDescriptor
  - shortNameOfInstitution
  - localEducationAgencyId
  - educationOrganizationIdentificationSystemDescriptor
  - identificationCode
- studentAssessments
  - studentAssessmentIdentifier
  - administrationDate
  - administrationEndDate
  - administrationLanguageDescriptor
  - whenAssessedGradeLevelDescriptor
  - assessmentIdentifier
  - namespace
  - schoolYear
  - studentUniqueId
  - performanceLevels.assessmentReportingMethodDescriptor
  - performanceLevelDescriptor
  - performanceLevelMet
  - scoreResults.assessmentReportingMethodDescriptor
  - scoreResults.resultDatatypeTypeDescriptor
  - scoreResults
- students
  - studentUniqueId
  - birthDate
  - firstName

    - lastSurname
    - middleName

- studentSchoolAssociations
    - entryDate
    - entryGradeLevelDescriptor
    - exitWithdrawDate
    - schoolReference.schoolId
    - schoolYearTypeReference.schoolYear
    - studentUniqueId

More information about these endpoints is available in the Ed-Fi Operational Data Store API.

2.3 Student Information

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

- Gender (M, F)
- Race
    - American Indian or Alaska Native
    - Asian
    - Black or African American
    - Hispanic or Latino
    - Native Hawaiian or Other Pacific Islander
    - Two or More Races
    - Unknown
    - White
- Economically Disadvantaged (Y, N) – only reported at aggregate levels
- English Learner (Y, N)
- Special Education (Y, N)
- Homeless (Y, N) – only reported at aggregate levels

2.4 Teacher Information

It is possible for Michigan educators to receive Teacher Growth reports from EVAAS. To provide these reports, SAS must receive teacher information from the MiDataHub to use in conjunction with MDE's student assessment data and the local interim/benchmark assessment data. This is necessary since the EVAAS models estimate the teacher growth measures for the group of students that are connected to a teacher in a given subject and grade. To receive this information, districts and/or ISDs must opt in to share MiDataHub data with SAS.

2.4.1 Data Used for Teacher-Student Linkages

The MiDataHub project contains different data tables for various purposes. EVAAS uses the following tables to obtain data for teacher-student linkages and/or identify the name and code of the district:

- schools
- students
- studentSectionAssociations
- staffSectionAssociations
- staffs
- courses
- courseOfferings
- academicSubjectDescriptors
- classroomPositionDescriptors
- CalendarDates
- Schools
- LocalEducationAgencies

The last three tables are only used to validate the district during the opt-in process.

Within these endpoints, SAS uses the following elements:

- academicSubjectDescriptors
  - codeValue
  - namespace
  - description/shortDescription
- classroomPositionDescriptors
  - codeValue
  - namespace
  - shortDescription
- courseOfferings
  - localCourseCode
  - localCourseTitle
  - courseCode
  - schoolId
  - schoolYear
  - sessionName
- courses
  - courseCode
  - courseTitle
  - academicSubjectDescriptor

- schools
    - schoolId
    - nameOfInstitution
    - operationalStatusDescriptor
    - schoolTypeDescriptor
    - shortNameOfInstitution
    - localEducationAgencyId
    - educationOrganizationIdentificationSystemDescriptor
    - identificationCode
- staffs
    - staffUniqueId
    - firstName
    - lastSurname
    - electronicMailAddress
    - electronicMailTypeDescriptor
    - staffIdentificationSystemDescriptor
    - identificationCode
- staffSectionAssociations
    - beginDate
    - endDate
    - classroomPositionDescriptor
    - teacherStudentDataLinkExclusion
    - localCourseCode
    - schoolId
    - schoolYear
    - sectionIdentifier
    - sessionName
    - staffUniqueId
- students
    - studentUniqueId
    - birthDate
    - firstName
    - lastSurname
    - middleName
- studentSectionAssociations
    - beginDate
    - endDate
    - teacherStudentDataLinkExclusion
    - localCourseCode
    - schoolId
    - schoolYear
    - sectionIdentifier

- sessionName
- studentUniqueId

2.4.2 Assigning Subject Areas

EVAAS uses the localCourseTitle element from the "courseOfferings" and "courses" endpoints to categorize courses as either Math or ELA for the state assessments. EVAAS also references the available subject area descriptions via the "AcademicSubjectDescriptors" endpoint and retains all records with subject areas relevant to our assessment pool (ELA/MATH).

Below are some examples of the values EVAAS looks for to identify the subject area that a course falls into. This list is not exhaustive.

- Mathematics: MATH, MTH, ALGEBRA, ALG, GEOMETRY
- English Language Arts: English, ELA, LANGUAGE ARTS, LANG ARTS, READ, LIT, L ARTS, LA

If EVAAS was not able to find teacher data for assessments for an entire grade in a school, the course names that were received for these students were reviewed. In these cases, coure names such as "Homeroom" or "GRADE 4" are categorized as both Math and ELA since these were most likely self-contained classrooms.

If courses were not categorized as either Math or ELA, corresponding records were dropped. Course names that were not categorized into either subject area were excluded. For example, course names that only referenced "Spelling," "Grammar," or "Writing" were not included.

2.4.3 Calculating Instructional Responsibility

EVAAS uses the student and teacher start and end dates to calculate how much of a student's instruction in a subject each teacher that interacted with that student is responsible for.

Percentages of instructional responsibility are based on two things:

1. The number of days a teacher taught a student in a tested subject compared to the total number of days the student was enrolled in the subject.
2. The number of teachers who taught the student.

Capturing the proportion of instructional responsibility for each teacher at the individual student level ensures EVAAS Teacher Value-Added reports link student growth to teachers fairly and accurately.

When calculating instructional responsibility, EVAAS uses the start and end dates for a student and teacher to determine how long a teacher provided instruction to a student in a course/subject. A student appears on a teacher's roster if the student's start and end dates overlap with the teachers.

If a student appears on multiple teachers' rosters for the same subject at the same time, instructional responsibility is split across the teachers. Here are two scenarios:

- Bobby is in a year-long grade 5 Math course, and Mrs. Smith is the only teacher of record for that course for all the days that Bobby is in that class. Mrs. Smith's instructional responsibility is 100%.
- Two teachers co-teach Bobby's year-long grade 5 Math course for the entire year. Each teacher has 50% instructional responsibility for Bobby.

In addition, EVAAS calculates the proportion of the school year each student received instruction in the tested subject. If a student was not enrolled for the entire school year, EVAAS adjusts the teacher's instructional responsibility to reflect the student's shortened instruction time. For example:

- Bobby's family moves to the area, and he enrolls in Mrs. Smith's year-long grade 5 Math course on the 45th day of the 180-day school calendar. Because he was in the class for 135 days and Mrs. Smith has 100% instructional responsibility, Mrs. Smith has 75% of the instructional responsibility for him.
- Bobby's family moves to the area, and he enrolls in Mrs. Smith and Mr. Jones' split year-long grade 5 Math course on the 45th day of the 180-day school calendar. Because he was in the class for 135 days and he splits his time with Mrs. Smith and Mr. Jones, each teacher has 37.5% of the instructional responsibility for him.
- Mrs. Smith was hired to replace Mr. Jones on the 90th day of the 180-day school calendar. Mrs. Smith and Mr. Jones have 50% instructional responsibility for the class.

Other business rules that affect the linkage data include:

- If a teacher's start and end dates are not populated, EVAAS uses the start and end dates for the student that the teacher is connected to.
- Courses that fall under the umbrella of Math or ELA are linked to the corresponding students' test scores. However, if a student is linked to both a general grade-level Math teacher and a Geometry teacher, EVAAS only links that student to the general grade-level Math teacher. If no general Math teacher exists, EVAAS links the student to the Geometry teacher.
- For courses that fall under the umbrella of Math or ELA, EVAAS attributes students' enrollment to their assessments. This means that the instructional responsibility for a student is split across multiple teachers if a student is enrolled in multiple ELA courses simultaneously.
- EVAAS excludes courses that are not discernable as Math or ELA from analysis. There are exceptions to this exclusion rule for districts, schools, and grades that have very low linkage rates or in cases where the courses table can be used to manually verify the subject area correctly indicates ELA or Mathematics.

2.4.4 Records Dropped in Initial Processing for Teacher-Student Linkages

There are several reasons why student and teacher data submitted through the tables in the MiDataHub might be removed through EVAAS' initial data processing. Some examples are listed below.

- No live data within MiDataHub at the time of the pull for the required endpoints.
- EVAAS connects data from the "students" endpoint and the "studentSectionAssociations" endpoint using UIC (studentUniqueId). If a UIC is present in one endpoint and not the other, then the record is incomplete and will be excluded.
- The course information provided in the "studentSectionAssociations" endpoint must have connecting course information present in the "staffSectionAssociations" endpoint so that a teacher record and student record can be connected. If the course information is present in one endpoint but not the other, then those records are excluded.
- If EVAAS is unable to identify these course titles referencing the "courseOfferings" or "courses" endpoint, then these courses are dropped. This connection does not exist, and the records are

excluded. The course information present in our "studentSectionAssociations" and "staffSectionAssociations" endpoints must have identifiable course name information within the "courseOfferings" or "courses" endpoint. If EVAAS is unable to identify these course with titles referencing the "courseOfferings" or "courses" endpoint, then these courses are dropped.

- If a value of studentUniqueID does not exist in any assessment data that EVAAS has received, then those records are excluded.
- If teacher and student dates do not overlap, then those records are removed from processing.

2.4.5 Combining Teacher-Student Linkages with Assessment Records

Once there is final set of teacher-student linkages, that information is connected to the assessment records to be used in the teacher value-added models. Students will have to meet other requirements described in the remainder of this document to be included in the teacher's growth measure.

3 Value-Added Analyses

The conceptual explanation of value-added reporting is the following:

- Growth = current achievement/current results compared to prior achievement/prior results with achievement being measured by a state summative assessment, such as M-STEP/PSAT 8.

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

- A **gain model** is used for tests given in consecutive grades, like the M-STEP Math and ELA in grades 3–7 to provide growth measures in grades 4–7 or MAP Math and Reading in grades 1–8 to provide growth measures in grades 1–8. The gain model is also used to measure growth from grade 7 to 8 with the PSAT 8/9 in grade 8. This model is known more formally as the Multivariate Response Model (or MRM).
- A **predictive model** is used for tests given in multiple grades or when performance from previous tests is used to predict performance on another test, such as the M-STEP Science and Social Studies assessments, SAT, and PSAT for grades 9 and 10. This model is known more formally as the Univariate Response Model (or URM).

Both models offer the following advantages:

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

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

Because EVAAS models use multiple subjects and grades for each student, it is typically not necessary to make *direct* adjustments for students' background characteristics. These adjustments are not necessary because each student serves as their own control. To the extent that socioeconomic and demographic influences persist over time, these influences are already represented in the student's data. As a 2004 study by The Education Trust stated, specifically with regard to the EVAAS modeling:

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

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

The value-added reporting based on statewide summative assessments in Michigan is available for districts, schools, and teachers. Teacher reporting is only available to those districts that have chosen to opt in through MiDataHub, and it can be based on either statewide summative assessments or the district's interim/benchmark assessments.

3.1 Gain Model

The gain model for districts, schools, and teachers is known as a multivariate response model, which can also be described as *linear mixed models* and *repeated measures models*. The district and school models are essentially the same, and the teacher model uses a slightly different approach that is more appropriate with the smaller numbers of students typically found in teachers' classrooms.

The gain-based model measures growth between two points in time for a group of students. The current growth expectation is met when a cohort of students moving from one grade to the next maintains the same relative position with respect to statewide student achievement in a specific subject. This relative position is specific to each year, subject, and grade. (See growth expectation in Section 4 for more details.)

For state summative assessments, growth is measured from one year to the next, using the available consecutive grade assessments. For MAP assessments, growth is measured from the beginning of the year to the end of the year within the same grade. Due to suspended assessments in the spring of the 2019-20 school year, the MAP assessments measure growth from the beginning of the year to the middle of the year within the same grade for the 2019-20 reporting.

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

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

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

- On the **choice of a complex value-added model**: McCaffrey, Daniel F., and J.R. Lockwood. 2008. "Value-Added Models: Analytic Issues." Prepared for the National Research Council and the National Academy of Education, Board on Testing and Accountability Workshop on Value-Added Modeling, Nov. 13-14, 2008, Washington, DC.

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

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

3.1.1 Gain Model at the Conceptual Level

As a simple example, consider the following scenario. Ten students are given a test in two different years. The goal is to measure academic growth (gain) from one year to the next. The right side of Figure 1 shows the same students, some of whom now have missing scores. Two simple approaches when data are missing are to calculate the mean of the differences or to calculate the differences of the means. When there is no missing data, these two simple methods provide the same answer (5.8 in the left-hand side of Figure 1). However, when there is missing data, each method provides a different result (6.9 versus 4.6 in the right-hand side of Figure 1).

Figure 1: Scores Without Missing Data, and Scores with Missing Data

Student	Previous Score	Current Score	Gain
1	51.9	74.8	22.9
2	37.9	46.5	8.6
3	55.9	61.3	5.4
4	52.7	47.0	-5.7
5	53.6	50.4	-3.2
6	23.0	35.9	12.9
7	78.6	77.8	-0.8
8	61.2	64.7	3.5
9	47.3	40.6	-6.7
10	37.8	58.9	21.1
Column Mean	**50.0**	**55.8**	**5.80**
Difference between Current and Previous Score Means			**5.80**

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

The gain model uses the correlation between current and previous scores in the nonmissing data to estimate a mean for the set of all previous and all current scores as if there were no missing data. It does this without explicitly assigning values for the missing scores. The difference between these two estimated means is an estimate of the average gain for this group of students. In this small example, the estimated difference on the right is 5.8 when using the gain model to first estimate the means in each column and taking the difference.

Even in a small example such as this, the estimated difference is much closer to the difference with no missing data on the left than either measure obtained by the mean of the differences (6.9) or difference of the means (4.6) on the right. This method of estimation has been shown, on average, to outperform both 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 Michigan, provide better correlation estimates by having more student data and more subjects and grades, which in turn provide better estimates of means and gains.

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

3.1.2 Normal Curve Equivalents

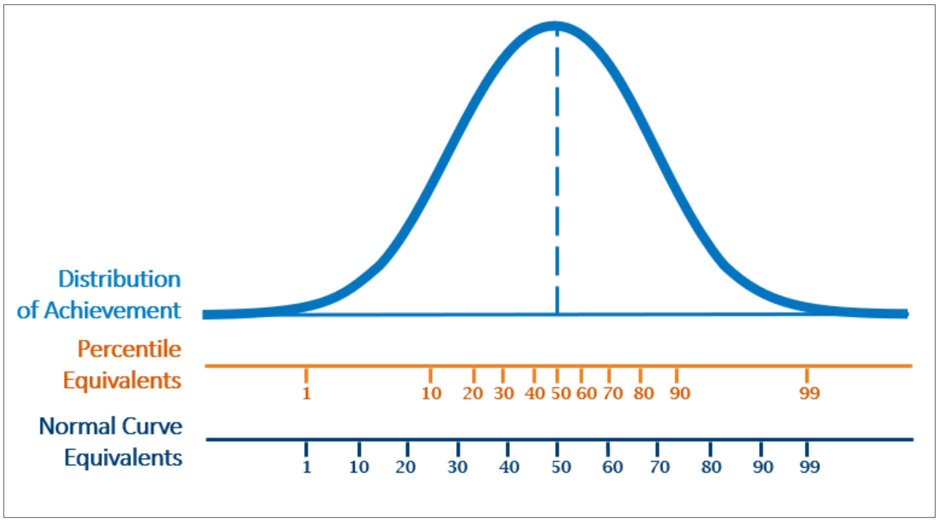
3.1.2.1 Why EVAAS Uses Normal Curve Equivalents in the Gain Model

The gain model estimates academic growth as a "gain," or the difference between two measures of achievement from one point in time to the next. For such a difference to be meaningful, the two measures of achievement (that is, the two tests whose means are being estimated) must measure academic achievement on a common scale. Some test companies supply vertically scaled tests 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 as seen in Figure 2 below. 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.

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

Figure 2: Normal Curve Equivalents



For example, in a typical year in Michigan, the average maximum NCE is approximately 108, corresponding to percentile rankings above 99.0. However, for display purposes in the EVAAS web application and to avoid confusion among users with interpretation, NCEs are shown as integers from 1-99. Truncating would create an artificial ceiling or floor, which might bias the results of the value-added measure for certain types of students. This forces the gain to be close to 0, or even negative, so the actual calculations use non-truncated numbers.

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

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

3.1.2.2 How EVAAS Uses Normal Curve Equivalents in the Gain Model

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

The first five columns of Table 1 show an example of a tabulated distribution of test scores from Michigan data. The tabulation shows, for each possible test score, in a particular subject, grade, and year, how many students made that score (“Frequency”) and what percentage (“Percent”) that frequency was out of the entire student population. (In Table 1, the total number of students is approximately 109,000). Also tabulated are the cumulative frequency (“Cum Freq,” which is the number of students who made that score or lower) and its associated percentage (“Cum Pct”).

The next step is to convert each score to a percentile rank, listed as "Ptile Rank" on the right side of Table 1. If a particular score has a percentile rank of 33, this is interpreted to mean that 33% of students in the population had a lower score and 67% had a higher score. In practice, there is some percentage of students that will receive each specific score. For example, 1.2% of students received a score of 1474 in Table 1. The usual convention is to consider half of that 1.2% to be "below" and half "above." Subtracting 0.6% (half of 1.2%) from the 33.6% who scored below the score of 1474 produces the percentile rank of 33.0 in Table 1.

Table 1: Converting Tabulated Test Scores to NCE Values

Score	Frequency	Cum Freq	Freq Pct	Cum Pct	Ptile Rank	Z	NCE
1474	1277	36632	1.2	33.6	33.0	-0.440	40.74
1475	1366	37998	1.3	34.8	34.2	-0.407	41.44
1476	1299	39297	1.2	36.0	35.4	-0.373	42.13
1477	1293	40590	1.2	37.2	36.6	-0.342	42.80
1478	1317	41907	1.2	38.4	37.8	-0.310	43.47
1479	1299	43206	1.2	39.6	39.0	-0.279	44.13
1480	1319	44525	1.2	40.8	40.2	-0.248	44.79

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

3.1.3 Technical Description of the Linear Mixed Model and the Gain Model

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

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

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

X is a known $m \times p$ matrix that allows inclusion of any fixed effects. Fixed effects are factors within the model that come from a finite population, such as all of the individual schools in the state of Michigan.

In the school model, there is a fixed effect for every school, year, subject, and grade. This matrix would have a row for each of these combinations.

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.1.3.1 District- and School-Level

The district and school gain models 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. Since gain models 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, which is also known as the R matrix. Specifically, scores belonging to the same student are correlated. If the scores in y are ordered so that scores belonging to the same student are adjacent to one another, then the R matrix is block diagonal with a block, R_i, for each student. Each student's R_i is a subset of the "generic" covariance matrix R_0 that contains a row and column for each subject and grade. Covariances among subjects and grades are assumed to be the same for all years (technically, all cohorts), but otherwise, the R_0 matrix is unstructured. Each student's R_i contains only those rows and columns from R_0 that

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

match the subjects and grades for which the student has test scores. In this way, the gain model can use all available scores from each student.

Algebraically, the district gain model is represented as:

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

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

The school gain model is represented as:

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

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

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

Solving the mixed model equations for the district or school gain model produces a vector b that contains the estimated mean score for each school (in the school model), subject, grade, and year. To obtain a value-added measure of average student growth, a series of computations can be done using the students from a school in a particular year and their prior and current testing data. The model produces means in each subject, grade, and year that can be used to calculate differences to obtain gains. Because students might change schools from one year to the next (in particular when transitioning from elementary to middle school, for example), the estimated mean score for the prior year and grade uses students that existed in the current year of that school. Therefore, mobility is considered within the model. Growth of students is computed using all students in each school including those that might have moved buildings from one year to the next.

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

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

3.1.3.2 Teacher-Level

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

- The current and previous teacher effects are incorporated.

- Each teacher estimate considers all the students' testing data over the years.
- The percentage of instructional responsibility (instructional time) the teacher has for each student is used.

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

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

Algebraically, the teacher model is represented as:

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

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 might have more than one teacher. The inner (rightmost) summation is over all the teachers of the i^{th} student in a particular subject/grade/year. $\tau_{ijk^*l^*t}$ is the effect of those teachers. $w_{ijk^*l^*t}$ is the fraction of the i^{th} student's instructional time claimed by the t^{th} teacher.

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

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

Table 2 illustrates how the Z matrix is encoded for three students who have three different scenarios of teachers during grades 3, 4, and 5 in two subjects, Math (M) and ELA (E).

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

Susan is taught by two different teachers in grade 3, teacher Abbot for Math and teacher Banks for ELA. In grade 4, Susan had teacher Card for ELA. For some reason, no teacher claimed Susan for Math in grade 4 even though she had a grade 4 Math test score. This score can still be included in the analysis by entering zeros into Susan's Z matrix rows for grade 4 Math. In grade 5, 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 5 ELA, there can be no corresponding row in Z.

Eric's scenario illustrates team teaching. In grade 3 Reading, Eric received an equal amount of instruction from both teachers Abbot and Banks. The entries in the Z matrix indicate each teacher's contribution, 0.5 for each teacher. In grade 5 Math, however, Eric was taught by both teachers East and Farr, but they did not make an equal contribution. Teacher East was attributed 80% instructional responsibility and teacher Farr was attributed 20% based on the fact both teachers taught the same students in the same course but for different lengths of time during the calendar year.

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

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

The teacher model provides estimated mean gains for each subject and grade. These quantities can be described by linear combinations of the fixed and random effects and are found using the equations mentioned above. Teachers receive separate estimates within each district since each individual district must opt-in through MiDataHub to provide data for use in the models.

Table 2: Encoding the Z Matrix

			Teachers											
			Third Grade				Fourth Grade				Fifth Grade			
			Abbot		Banks		Card		Dupont		East		Farr	
Student	**Grade**	**Subjects**	**M**	**E**	**M**	**E**	**M**	**E**	**M**	**E**	**M**	**E**	**M**	**E**
Tommy	**3**	**M**	1	0	0	0	0	0	0	0	0	0	0	0
		E	0	1	0	0	0	0	0	0	0	0	0	0
	4	**M**	1	0	0	0	1	0	0	0	0	0	0	0
		E	0	1	0	0	0	1	0	0	0	0	0	0
	5	**M**	1	0	0	0	1	0	0	0	1	0	0	0
		E	0	1	0	0	0	1	0	0	0	1	0	0
Susan	**3**	**M**	1	0	0	0	0	0	0	0	0	0	0	0
		E	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
		E	0	0	0	1	0	1	0	0	0	0	0	0
	5	**M**	1	0	0	0	0	0	0	0	0	0	1	0
Eric	**3**	**M**	1	0	0	0	0	0	0	0	0	0	0	0
		E	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
		E	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
		E	0	0.5	0	0.5	0	0	0	1	0	1	0	0

3.1.4 Where the Gain Model is Used in Michigan

The gain model is used with the M-STEP assessment in Math and in ELA for grades 3–7 to provide value-added measures at the district, school, and teacher level in grades 4–7 in Math and ELA for districts that have opted-in through MiDataHub. The gain model is also used to measure growth from grade 7 to 8 with the PSAT 8/9 in grade 8. Finally, it is also used with the MAP assessment in Math and Reading for grades 1–8 to provide value-added measures at the teacher level in grades 1–8 for districts that have opted in through MiDataHub.

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

3.1.5 Students Included in the Analysis

All students' scores are included in these analyses if the scores can be used and do not meet any criteria for exclusion outlined in Section 8. In other words, students' available Math and ELA results for the student's cohort are incorporated in the state summative models, and the students' available MAP Math and Reading results for the student's cohort are incorporated in the interim/benchmark models. In the gain model for state summative assessments, students' scores that do not meet Full Academic Year are excluded from all years in the analyses.

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

3.1.6 Minimum Number of Students for Reporting

3.1.6.1 District- and School-Level

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

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 seven students who are associated with the school or district in that subject/grade/year.
- There is at least one student at the school or district who has a "simple gain," which is based on a valid test score in the current year/grade as well as the prior year/grade in the same subject.
- Of those students who are associated with the school or district in the current year/grade, at least seven students must be in the prior year/subject/grade to generate a gain in the current year/subject/grade.

3.1.6.2 Teacher-Level

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

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

As the second requirement, the teacher must be linked to at least seven students with prior test score data in the same subject, and the test data might come from any prior grade (or beginning of year semester for MAP) if they are part of the student's regular cohort. (If a student repeats a grade, then the prior test data would not apply as the student has started a new cohort.) One of these seven students must have a "simple gain," meaning the same subject prior test score must come from the immediate prior year and prior grade for state assessments or the beginning of year semester of the current year and grade for MAP assessments. Students are linked to a teacher based on the subject area taught and the assessment taken. Students that have no prior testing data in the same subject area are not linked to the teacher for the analysis.

3.2 Predictive Model

Tests that are not necessarily administered to students in consecutive years, like M-STEP Science and Social Studies, require a different modeling approach from the gain-based model. For these tests, EVAAS reporting uses a predictive model called the univariate response model (URM). This model is also used when previous test performance is used to predict another test's performance, such as the SAT. 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 predictive model 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, or teacher made the same amount of growth as students in the average district/school/teacher with the state for that same year, subject, and grade. Teacher reporting is not currently used in any subjects and grades that use the predictive model.

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

- The model does not require students to have all predictors or the same set of predictors if a student has at least three prior test scores in any subject/grade.
- The model minimizes the influence of measurement error by using many prior tests for an individual student. Analyzing all subjects simultaneously increases the precision of the estimates.

- The model uses scores from multiple tests, including those on different scales.
- The model accommodates teaching scenarios where more than one teacher has responsibility for a student's learning in a specific subject/grade/year.

In Michigan, the predictive model is used for value-added reporting for M-STEP Science (when not in field testing) and social studies as well as PSAT 8/9 in grade 9, PSAT 10, and SAT. These PSAT- and SAT-based reports are available for districts and schools, not teachers, since these tests typically cover multiple content areas at the high school level.

3.2.1 Predictive Model at the Conceptual Level

The predictive model is run for each individual year, subject, and grade (if relevant). Consider all students who took M-STEP Social Studies for grade 8 in a given year. Those students are connected to their prior testing history (across grades, subjects, and years), and the relationship between the observed M-STEP Social Studies scores for grade 8 with all prior test scores is examined. It is important to note that some prior test scores are going to have a greater relationship to the score in question than others. For example, it might be that prior Social Studies or Reading tests will have a greater relationship with M-STEP Social Studies for grade 8 than prior Math scores. However, the other scores still have a statistical relationship.

Once that relationship has been defined, a predicted score can be calculated for each individual student based on their own prior testing history. With each predicted score based on a student's prior testing history, this information can be aggregated to districts or schools. 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 or school. 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 or school in terms of the average growth, so that if every student obtained their predicted score, a district or school 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 reference group) that year.

3.2.2 Technical Description of the District and School Models

The predictive model has similar models for districts and schools and a slightly different model for teachers that allows multiple teachers to share instructional responsibility. These models can be used at the teacher level, but they are not currently being used in Michigan in this way. The predictive model approach is described briefly below with more details following.

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

Algebraically, the model can be represented as follows for the i^{th} student.

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

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

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

Two difficulties must be addressed 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-school 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:

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

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

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

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

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

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

The $\hat{y}_i$ term is nothing more than a composite of all the student's past scores. It is a one-number summary of the student's level of achievement prior to the current year. The different prior test scores making up this composite are given different weights (by the regression coefficients, the $\hat{\beta}$s) to maximize its correlation with the response variable. Thus, a different composite would be used when the response variable is math than when it is reading for example. Note that the $\hat{\alpha}_j$ term is not included in the equation. Again, this is because $\hat{y}_i$ represents prior achievement before the effect of the current district or school. 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 predictive model is to estimate the school effects (α_j) using the following ANCOVA model:

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

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

3.2.3 Students Included in the Analysis

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

For the M-STEP and SAT analyses, students' scores that do not meet Full Academic Year membership in the current year are excluded from the analyses.

There are no membership rules used to include or exclude students in the PSAT 8/9 in grade 9 and PSAT 10 analyses. EVAAS does not provide teacher-level measures using general achievement tests like the College Board assessments since individual teachers are typically connected to specific courses such as Algebra I or Geometry. The information that is covered on college readiness-type assessments might be covered in multiple courses across years. Therefore, it is not recommended to use them for individual teacher-level growth measures.

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

3.2.4 Minimum Number of Students for Reporting

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

4 Growth Expectation

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

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

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

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

4.1 Description

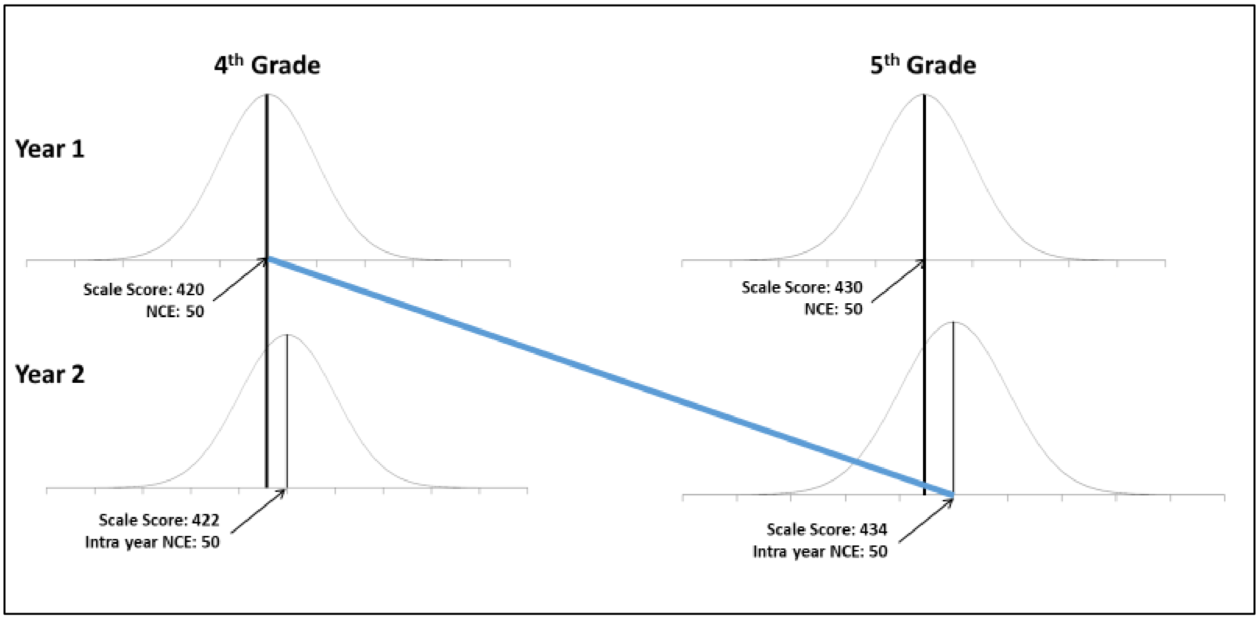
- The actual definitions in each model are slightly different, but the concept can be considered as the average amount of growth seen across the state in a statewide implementation for statewide assessments. For MAP, the concept can be considered as the average amount of growth observed in the national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

4.2 Illustrated Example

Figure 3 below provides a simplified example of how growth is calculated with this approach when the state achievement increases. The figure has four graphs, each of which plot the NCE distribution of scale scores for a given year and grade. The scale scores are used to illustrate an example in the graphics below and do not represent actual scale scores in Michigan. In this example, the figure shows how the gain is calculated for a group of grade 4 students in Year 1 as they become grade 5 students in Year 2. In Year 1, our grade 4 students score, on average, 420 scale score points on the test, which corresponds to the 50th NCE (similar to the 50th percentile). In Year 2, the students score, on average, 434 scale score points on the test, which corresponds to a 50th NCE *based on the grade 5 distribution of scores in Year 2*. The grade 5 distribution of scale scores in Year 2 was higher than the grade 5 distribution of scale scores in Year 1, which is why the lower right-hand graph is shifted slightly to the right. The blue line shows what is required for students to make expected growth, which would be to maintain their position at the 50th NCE in grade 4 in Year 1 as they become grade 5 students in Year 2. The growth measure for these students is Year 2 NCE – Year 1 NCE, which would be 50 – 50 = 0. Similarly, if a group of students started at the 35th NCE, the expectation is that they would maintain that 35th NCE.

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

Figure 3: Growth Expectation Example



4.3 Defining the Expectation of Growth During an Assessment Change

During the change of assessments, the scales from one year to the next will be completely different from one another. This does not present any particular changes with the predictive model 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. There will be a need for the predictors to be adequately related to the response variable of the new assessment, but that typically is not an issue.

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

Over the past 25 years, EVAAS reporting has accommodated several changes in testing regimes and used several tests for the gain model without a break in reporting.

5 Categorizing Growth Measures

EVAAS models provide more than just the value-added estimate (growth measure); other metrics are available such as the growth measure's associated standard error and the student-level standard deviation of growth for each year, subject, and grade. This section provides more information about how all these metrics are used to classify growth measures into meaningful categories for interpretation.

5.1 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 certainty for the growth estimate. This metric depends on the quantity and quality of student data included in the estimate, such as the number of students and the occurrence of missing data for those students. Taken together, the estimate and standard error provide educators and policymakers with critical information about the certainty that students in a district, school, or classroom are making decidedly more or less than the expected growth. Taking the standard error into account is particularly important for reducing the risk of misclassification.

Furthermore, because the gain and predictive models use robust statistical approaches as well as maximize the use of students' testing history, they can provide value-added estimates for relatively small numbers of students. This allows more teachers, schools, and districts to receive their own value-added estimates, which is particularly useful to rural communities or small schools. As described in Section 3, there are minimum requirements of students per tested subject/grade/year depending on the model, which are relatively small. The use of standard error ensures that there is sufficient evidence about students' growth prior to classifying teachers, schools, and districts, even when there are relatively small numbers of students involved.

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

5.2 Determining Growth Effect Size Using the Student-Level Standard Deviation of Growth

The student-level standard deviation of growth can be used to provide context about the magnitude of growth being made by a group of students. For the gain-based model (where this metric is applied), students typically have a current and a prior year NCE, which can be used to derive a student-level gain. The standard deviation of the student-level distribution of growth is available for each year, subject, and grade. Dividing the growth measures by the standard deviation provides a value known as an "effect size," and it indicates the practical significance regarding the group of students and whether they met, exceeded, or fell short of expected growth.

5.3 Categorizing District and School Growth Measures

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

The chart below provides the color-coding, definitions, and interpretation for the value-added reports of districts and schools.

Value-Added Color	Growth Measure Compared to the Growth Standard	Index	Interpretation
DB	At least 2 standard errors above	2.00 or greater	Significant evidence that the school's students made more progress than the growth standard
LB	Between 1 and 2 standard errors above	Between 1.00 and 2.00	Moderate evidence that the school's students made more progress than the growth standard
G	Between 1 standard error above and 1 standard error below	Between -1.00 and 1.00	Evidence that the school's students made progress similar to the growth standard
Y	Between 1 and 2 standard errors below	Between -2.00 and -1.00	Moderate evidence that the school's students made less progress than the growth standard
LR	More than 2 standard errors below	Less than -2.00	Significant evidence that the school's students made less progress than the growth standard

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

5.4 Categorizing Teacher Growth Measures

Teacher reporting will categorize teacher growth measures using a two-step process based on, first, the growth index and, second, the effect size.

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

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

For the first step with uncertainty, the thresholds are an index of +2 or greater, an index of -2 or less, or an index between -2 and +2. These thresholds are similar to the concept of a 95% confidence interval. If a 95% confidence interval around the growth measure did not contain the growth expectation, then they would fall outside the thresholds. The second step uses an effect size threshold of 0.4 and -0.4. These values correspond to a "medium" effect size as referenced in John Hattie's work.[4]

In accordance with MDE policies, there are four categories for teacher growth categorization. The top category has a growth index of greater than or equal to 2 *and* an effect size of greater than or equal to 0.4. The next highest category consists of all other measures where the growth index is greater than or equal to -2 *and* one other condition is met: either the index is also less than 2 *or* the effect size is less than 0.4. The bottom category is when the growth index is less than -2 and the effect size is less than -0.4. The next to bottom category are teachers with a growth index less than -2, but their effect size is greater than or equal to -0.4.

The chart below provides the color-coding, definitions and interpretation for the Value-Added reports of *teachers*.

Value-Added Color	**Definition**	**Interpretation**
DB	Index is greater than or equal to 2 *and* the effect size is greater than or equal to 0.40	Level 4, Exceeds: Significant evidence that the teacher's students made more progress than the growth standard and the effect size is medium or higher
G	Index is greater than or equal to -2 *and* either the index is less than 2 or the effect size is less than 0.4.	Level 3, Met: Evidence that the teacher's students made progress similar to the growth standard
Y	Index is less than -2 *and* the effect size is greater than or equal to -0.4.	Level 2, Nearly Met: Significant evidence that the teacher's students made less progress than the growth standard but not less than a negative medium effect size
LR	Index is less than -2 *and* the effect size is less than -0.4.	Level 1, Not Met: Significant evidence that the teacher's students made less progress than the growth standard and less growth than a negative medium effect size

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

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

[4] See, for example, John Hattie, *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement* (London: Routledge, 2008).

5.5 Rounding and Truncating Rules

As described in Section 5.4, the effectiveness categories are based on the value of the growth index. In determining the growth index, rounding and truncating rules are applied only in the final step of the calculation. Thus, the calculation of the growth index uses unrounded values for the value-added measures and standard errors. After the growth index has been created but before the categories are determined, the index values are rounded or truncated by taking the maximum value of the rounded or truncated index value out to two decimal places. This business rule yields the highest category of effectiveness given any type of rounding or truncating situation. For example, if the index score was a 1.995, then rounding would provide a higher category. If the score was a -2.005, then truncating would provide a higher category. In practical terms, this only impacts a small number of measures. The same rules are also applied to effect sizes for teachers.

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

6 Composite Calculations

6.1 Teacher Composites

6.1.1 Available Teacher Composites

EVAAS reporting provides growth measures by test, subject, grade, and year, but these measures can be combined across subjects, grades and/or years. Teacher reporting is available for the M-STEP Math and ELA assessments and PSAT 8/9 in grade 8 as well as MAP Math and Reading in grades 1–8. It is not available for the other PSAT and SAT assessments because they are not course-specific and tend to assess general content knowledge that would be covered in several courses at the high school level. Teacher reporting is available for the 2017-18 and 2018-19 school years for state summative assessments and for the 2017-18, 2018-19, and 2019-20 (MOY) school years for MAP.

Teachers will received a composite if they have teacher reporting available for the most recent year of reporting (2018-19 for state summative assessments and 2019-2020 [MOY] for MAP assessments). Depending on what is available for the teacher, the following composites are available:

- Subject-specific composite across grades for a given type of test, such as:
  - Up to three year M-STEP Math for grades 4–7 and PSAT 8/9 in grade 8
  - Up to three year M-STEP ELA for grades 4-7 and PSAT 8/9 in grade 8
  - Up to three year MAP Math for grades 1–8
  - Up to three year MAP Reading for grades 1–8
- Overall composite across subjects and grades for a given type of test, such as:
  - Up to three year M-STEP Math and ELA for grades 4–7 and PSAT 8/9 in grade 8
  - Up to three year MAP Math and Reading for grades 1–8

Note that these composites are based on one type of assessment, state summative *or* MAP, not both combined. Based on MDE policy, these composites will include up-to-three years of growth data. If a teacher only has one year of growth data for the most recent year, then that teacher's composite only includes growth data from that single year.

6.1.2 Business Rules for Combining Growth Measures into Teacher Composites

The following business rules apply to teacher composites in accordance with MDE policies and preferences.

- A composite is weighted by the number of "full-time equivalent" students associated with each individual growth measure for the type of assessment (state summative or MAP).
- For each teacher, the full-time equivalent (FTE) number of students is based on the number of students linked to that teacher as well as the percentage of instructional time the teacher has for each student. For example, if a teacher taught 26 students for 50% of their instructional time, then the teacher's student FTE number would be 26 students times 50% of their instructional learning time, or 13 students.
- Typically this growth is combined within a year first and then across years.
- The across-year measures are also weighted by the student FTE number.

6.1.3 Calculation of Teacher Composites

Composite calculations vary according to what is included in the composite in terms of tests, subjects, grades, and years for a given teacher.

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

1. Calculate the *gain* across grades and subjects for a given year.
2. Calculate the *standard error* across grades and subjects for a given year.
3. Calculate the composite *gain* across years.
4. Calcuate the composite *standard error* across years.
5. Calculate the composite *index* across years.
6. Calculate the composite *effect size* across years.

The following sections illustrate this process of creating the across-subject composite for a teacher who has two years of data.

Table 3: Sample Teacher Value-Added Information

Year	Subject	Grade	Growth Measure	Standard Error	Index	Std. Dev.	Effect Size	Number of FTE Students
2019	Math	6	3.30	0.70	4.71	11.0	0.30	25
2019	ELA	6	-1.10	1.00	-1.10	10.0	-0.11	23
2018	Math	6	1.70	0.65	2.62	10.5	0.16	27

6.1.4 Calculate Gain Across Grades and Subjects for a Given Year

Because all growth measures from the gain-based model are in the same scale (Normal Curve Equivalents), the teacher composite gain across the two applicable subject/grades is a weighted average of the individual gains based on the number of effective students in each subject and grade. For the teacher, the total number of FTE students affiliated with gain-based growth measures in 2019 is 25 + 23, or 48. The 2019 grade 6 Math value-added measure would be weighted at 25/48, the 2019 grade 6 ELA value-added measure would be weighted at 23/48. More specifically, the composite gain is calculated using the following formula:

$$2019\ Comp\ Gain = \frac{25}{48}Math_6 + \frac{23}{48}ELA_6 = \frac{25}{48}(3.30) + \frac{23}{48}(-1.10) = 1.19 \quad (20)$$

6.1.5 Calculate Standard Error Across Grades and Subjects for a Given Year

6.1.5.1 Technical Background on Standard Errors

The standard error of the gain-based teacher composite gain cannot be calculated using the assumption that the gains making up the composite are independent. This is because many of the same students are likely represented in different value-added gains, such as grade 6 Math in 2018 and grade 6 ELA in 2018. The statistical approach, outlined in Section 3.1.3 (with references), is quite sophisticated and will consider the correlations between pairs of value-added gains as shown in equation (21) below and using equation (10) for teachers.[5] The composites are indeed linear combinations of the fixed effects of the models and can be estimated as described in Section 3.1.3. The magnitude of each correlation depends on the extent to which the same students are in both estimates for any two subject, grade, and year estimates.

6.1.5.2 Illustration of Gain-Based Standard Error for Sample Teacher

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

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

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

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

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

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

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

Two variables that are unrelated have a correlation and covariance of zero. Such variables are said to be statistically independent. If the X and Y values have a positive relationship, then the covariance will also be positive. As a general rule, two value-added gain estimates are statistically independent if they are based on completely different sets of students.

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

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

student weightings and standard errors reported in Table 3 and assuming total independence, the standard error would then be as follows:

$$\begin{aligned} 2018\ Comp\ SE &= \sqrt{\left(\frac{25}{48}\right)^2 (SE\ Math_6)^2 + \left(\frac{23}{48}\right)^2 (SE\ ELA_6)^2} \\ &= \sqrt{\left(\frac{25}{48}\right)^2 (0.70)^2 + \left(\frac{23}{48}\right)^2 (1.00)^2 = 0.60} \end{aligned} \tag{23}$$

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

In this example, since the teacher teaches the same grade in different subjects, the actual standard error will likely be above the value of 0.60 due to students being in both Math and ELA with the teacher. The specific value will depend on the values of the correlations across the two gains. Correlations of gains across years might be positive or slightly negative since the same student's score can be used in multiple gains if a teacher has taught that student multiple times. The magnitude of each correlation depends on the extent to which the same students are in both estimates for any two subject/grade/year estimates.

For the sake of simplicity, let us assume the actual standard error was 0.65 for the teacher composite in this example.

6.1.6 Calculate Composite Gain Across Years

The next step is to calculate the gain for students across time for this teacher. The composite gain would be found by taking the weighted average of year's gain as follows:

$$Comp\ gain = \frac{48}{75} gain_{2019} + \frac{27}{75} gain_{2018} = \frac{48}{75}(1.19) + \frac{27}{75}(1.70) = 1.37 \tag{24}$$

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

6.1.7 Calculate Composite Standard Error Across Years

The calculations above provide the composite gain across years. Then we have a standard error for each year. These can be combined to create a standard error for the composite gain. Assuming independence across time and using the student weightings and single-year standard errors, the multi-year standard error would then be as follows:

$$Comp\ SE = \sqrt{\left(\frac{48}{75}\right)^2 (SE_{2019})^2 + \left(\frac{27}{75}\right)^2 (SE_{2018})^2} = \sqrt{\left(\frac{48}{75}\right)^2 (0.60)^2 + \left(\frac{27}{75}\right)^2 (0.65)^2} = 0.45$$

6.1.8 Calculate Composite Index Across Years

The next step is to calculate the teacher composite index, which is the teacher composite value-added gain divided by its standard error. The gain-based composite index for this teacher would be calculated as follows:

$$Comp\ Index = \frac{Comp\ Gain}{Comp\ SE} = \frac{1.37}{0.45} = 3.04 \tag{24}$$

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

6.1.9 Calculate the Composite Effect Size Across Years

To calculate the effect size for the overall composite, each growth measure is divided by the student-level standard deviation of growth. This value is a constant within each year subject and grade but can be different across the different year, subject, and grades. The composite effect size is a weighted average of the effect sizes based on the FTE number of students.

$$\begin{aligned} Comp\ Effect\ Size &= \frac{25}{75}Math_{2019_6} + \frac{23}{75}ELA_{2019_6} + \frac{27}{75}Math_{2018_6} \\ &= \frac{25}{75}(0.30) + \frac{23}{75}(-0.11) + \frac{27}{75}(0.16) = 0.12 \end{aligned} \tag{25}$$

6.1.10 Categorizing Growth Measures as a Final Step

With the combined composite growth index and effect size, the specific composite can be categorized. This growth index is above 2.00. The effect size is below 0.40. Therefore, based on Section 5.4, the teacher composite would fall into Level 3 or Met.

7 Projection Model

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

Currently, the following projections are available to educators in Michigan:

- M-STEP Math and ELA in grades 5–7
- M-STEP Social Studies in grades 5, 8, and 11
- M-STEP Science in grades 5, 8, and 11 (not available in 2019 due to field testing)
- PSAT 8/9 Mathematics and ELA in grade 8
- PSAT 8/9 in Mathematics, Evidence-Based Reading and Writing in grade 9
- PSAT 10 Mathematics and Evidence-Based Reading and Writing in grade 10
- SAT Mathematics and Evidence-Based Reading and Writing

The 2018-19 reporting includes M-STEP and PSAT 8/9 projections to one or two grades above the last tested grade. 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 described in Section 3.2.2. In this model, the projected score serves as the response variable (y), the covariates (xs) are scores on tests the student has already taken, and the categorical variable is the school at which the student received instruction in the subject, grade, and year of the response variable (y). Algebraically, the model can be represented as follows for the i^{th} student.

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i \tag{26}$$

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

$$\hat{y}_i = \hat{\mu}_y \pm \hat{\alpha}_j + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots + \epsilon_i \tag{27}$$

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

Two difficulties must be addressed to implement the projections. First, not all students will have the same set of predictor variables due to missing test scores. Second, because 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), such as Proficient or Advanced on a future M-STEP. Examples of these benchmarks include the following:

- Probability that last tested fourth graders score Partially Proficient, Proficient, or Advanced on the fifth-grade M-STEP assessment
- Probability that last tested fifth graders score Partially Proficient, Proficient, or Advanced on the sixth-grade M-STEP assessment
- Probability that last tested sixth graders score Partially Proficient, Proficient, or Advanced on the seventh-grade M-STEP assessment
- Probability that last tested seventh graders score Partially Proficient, Proficient, or Advanced on the PSAT 8/9 in grade 8
- Probability that last tested eighth graders meet the ninth-grade standard on the PSAT 8/9 in grade 9
- Probability that last tested ninth graders meet the standard on the PSAT 10
- Probability that last tested ninth and 10th graders score Partially Proficient, Proficient, or Advanced on the SAT

The probability is calculated as the area above the benchmark cutoff score using a normal distribution with its mean equal to the projected score and its standard deviation equal to the standard error of the projected score as described below. Φ represents the standard normal cumulative distribution function.

$$Prob(\hat{y}_i \geq b) = \; \Phi\left(\frac{\hat{y}_i - b}{SE(\hat{y}_i)}\right) \qquad (28)$$

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

8.2 Checks of Scaled Score Distributions

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

8.2.1 Stretch

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

8.2.2 Relevance

Relevance indicates whether the test has sufficient alignment with the state standards for statewide assessments or with local standards for the interim/benchmark assessments. The requirement that tested material will correlate with standards if the assessments are designed to assess what students are expected to know and be able to do at each grade level. This is how state tests are designed and is monitored by the MDE and their psychometricians. More details about MDE-approved benchmark assessment providers are available here: https://www.michigan.gov/mde/0,4615,7-140-22709_102327---,00.html.

8.2.3 Reliability

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

8.3 Data Quality Business Rules

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

8.3.1 Missing Grade Levels

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

8.3.2 Duplicate (Same) Scores

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

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

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

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

If a student has multiple scores in the same period for a particular subject and grade and the test scores are not the same, then those scores will be excluded from the analysis. For MAP assessments, if a student has multiple scores in the same period, the BOY test score is defined as the first test date for a student/test/subject/grade and the MOY and EOY test scores are defined as the last test date for a student/test/subject/grade. 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. This is applied to state assessments and any remaining MAP assessment records that could not be resolved by the first and last test date business rule.

If duplicate scores for a particular subject and tested grade in a given testing period are at different schools, then both scores will be excluded from the analysis.

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 whether the data for two separate students were inadvertently combined. If this is the case, then student data are adjusted so that each unique student is associated with only the appropriate scores. If all the scores appear to be associated with a single unique student, then scores that appear inconsistent are excluded from the analysis.

8.3.6 Students with Records That Have Unexpected Grade Level Changes

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

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

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

8.3.8 Outliers

8.3.8.1 Conceptual Explanation

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

8.3.8.2 Technical Explanation

Student assessment scores are checked each year to determine whether they are outliers in context with the other scores in a reference group of scores from the individual student. These reference scores are weighted differently depending on proximity in time to the score in question. Scores are checked for outliers using related subjects as the reference group. For example, when searching for outliers for Math test scores on state assessments, all Math scores from state assessments are examined simultaneously during outlier identification for the state assessments, and any scores that appear inconsistent, given the other scores for the student, are flagged. Outlier identification for college readiness assessments use all available college readiness data alongside state assessments in the respective subject area (e.g., Math subjects with M-STEP and PSAT tests might be used to identify outliers with SAT). Scores are flagged in a conservative way to avoid excluding any student scores that should not be excluded. Scores can be flagged as either high or low outliers. Once an outlier is discovered, that outlier will not be used in the analysis, but it will be displayed on the student testing history on EVAAS web application.

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

Considerations included in outlier detection are:

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

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

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

For low-end outliers, the rules are:

- The percentile of the score must be below 50.
- The t-value must be below -3.5 for M-STEP grades 3–7 and PSAT 8/9 in eighth grade for Math and ELA and for MAP grades 1–8 for Math and Reading when determining the difference between the score in question and the weighted combination of reference scores (otherwise known as the comparison score). In other words, the score in question must be at least 3.5 standard deviations below the comparison score. For other assessments, the t-value must be below -4.0.
- The percentile of the comparison score must be above a certain value. This value depends on the position of the individual score in question but will need to be at least 10 to 40 percentiles above the individual percentile score.

For high-end outliers, the rules are:

- The percentile of the score must be above 50.
- The t-value must be above 4.5 for M-STEP grades 3–7 and PSAT 8/9 in eighth grade for Math and ELA and for MAP grades 1–8 for Math and Reading when determining the difference between the score in question and the reference group of scores. In other words, the score in question must be at least 4.5 standard deviations above the comparison score. For other assessments, the t-value must be above 5.0.
- The percentile of the comparison score must be below a certain value. This value depends on the position of the individual score in question but will need to be at least 30 to 50 percentiles below the individual percentile score. There must be at least three reference scores used to make the comparison score.