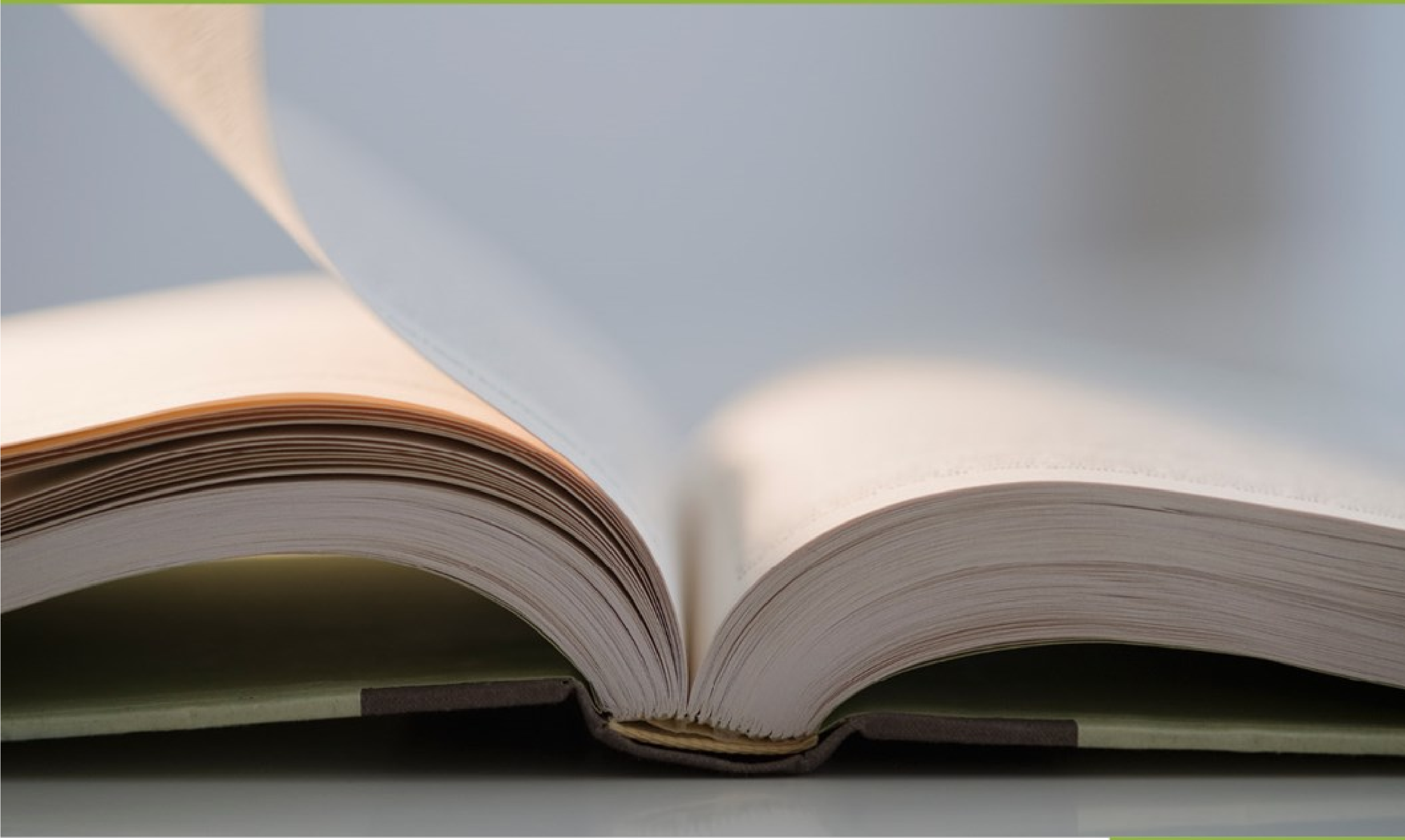
SAS® EVAAS

Statistical Models and Business Rules

Prepared for the Iowa Department of Education



THE POWER TO KNOW.

Contents

1 Introduction to Iowa’s Growth Reporting

EVAAS growth models use established statistical analysis to measure students’ academic growth. Conceptually, growth measures are calculated by comparing the exiting achievement to the entering achievement for a group of students. Although the concept of growth is easy to understand, the implementation of a growth model is more complex.

First, there is not just one way to measure growth. There are multiple growth models that vary based on the intended use of the data, students included in the analysis, and level of reporting (district, school, or classroom). To better understand the differences between models, this document provides detail about the business rules used for calculating each growth measures that reflect the policies and practices selected by the Iowa Department of Education.

Second, in order to provide reliable growth measures, growth models must overcome the complexities of working with student assessment data. A few examples are students who do not have the same entering achievement, students who do not have the same set of prior test scores, and all assessments have measurement error because they are estimates of student knowledge.

Third, the growth measures are relative to students’ expected growth. Expected growth is determined by the growth that is observed within the actual population of Iowa test-takers in a subject, grade, and year. Examining the difference between a student’s current achievement and their expected achievement creates a growth measure that is more nuanced and statistically robust.

With these complexities in mind, the purpose of this document is to provide information about the EVAAS growth model. This document includes information about the statistical models, business rules, policies, and practices implemented in the reports available in the EVAAS platform.

Specifically, more information can be found in these areas:

- Conceptual and technical explanations of analytic models
- Definition of expected growth
- Classifying growth into categories for interpretation
- Input data
- Business rules

Although the underlying statistical models and business rules supporting these reports are sophisticated and comprehensive, the web reports are designed to be user-friendly so that educators and administrators can quickly identify strengths and opportunities for improvement. They then can use these insights to inform curricular, instructional, and planning supports.

2 Statistical Models

2.1 Overview of Statistical Models

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

The EVAAS value-added models offer the following advantages:

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

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

Furthermore, because this robust modeling approach uses multiple years of test scores for each student and includes students who are missing test scores, EVAAS growth measures typically have very low correlations with student characteristics. It is not necessary to make *direct* adjustments for student socioeconomic status or demographic flags because each student serves as their own control. In other words, to the extent that background influences persist over time, these influences are already represented in the student's data. As a 2004 study by The Education Trust states, 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.[1]

While technically feasible, the value-added reporting in Iowa does not make any direct adjustments for students' socioeconomic or demographic characteristics because the students' results are already

[1] 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.

influenced by these characteristics. This makes including additional adjustments in the statistical models unnecessary. Through this approach, the Iowa Department of Education does not promote a growth model with differential expectations for groups of students based on their backgrounds.

EVAAS includes reports displaying results from the following models within the web application:

- **The Predictive model** is used to measure growth among groups of students on a particular assessment by comparing students' expected performance to their actual performance.
- **The Projection model** is used to estimate the probability of obtaining a particular score or higher on a future assessment for individual students. This is similar to the predictive model except that it is intended as an instructional tool for educators serving students who have not yet taken an assessment.

The following sections provide technical explanations of the models. The online Help within the EVAAS web application provides educator-focused descriptions of the models.

2.2 Predictive Model

2.2.1 Overview

The predictive model is a regression-based model where growth is a function of the difference between students' expected scores and their actual scores. Expected growth is met when students within a district, school, or classroom made the same amount of growth as students within the average district, school, or classroom.

There are three separate analyses for EVAAS reporting based on the predictive model: one each for districts, schools, and classrooms. The district and school models are essentially the same. In contrast, the classroom model includes accommodations for team teaching and other shared instruction.

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

2.2.2 Conceptual Explanation

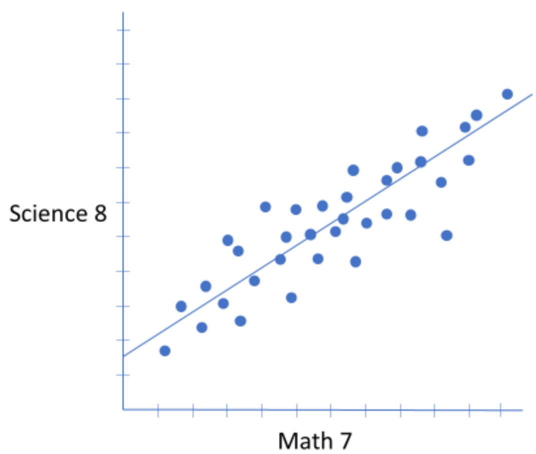
As mentioned above, the predictive model uses a student's previous test performance to create a prediction of their test performance. Growth is measured by comparing this performance expectation to actual performance. This model is suitable for tests given in consecutive grades, such as Math and English Language Arts (ELA) assessments in grades 3–11, but it can also handle scenarios where tests are not given consecutively, such as Science assessments in grades 5, 8, and 10. The predictive model is able to handle these scenarios because it is based on students' prior testing histories. This leads to several advantages:

- It minimizes the influence of measurement error and increases the precision of predictions by using multiple prior test scores as predictors for each student.
- It does not require students to have all predictors or the same set of predictors as long as a student has at least three prior test scores as predictors of the response variable in any subject and grade.

- It allows educators to benefit from all tests, even when tests are on different scales.
- It accommodates teaching scenarios where more than one teacher has responsibility for a student's learning in a specific subject, grade, and year.

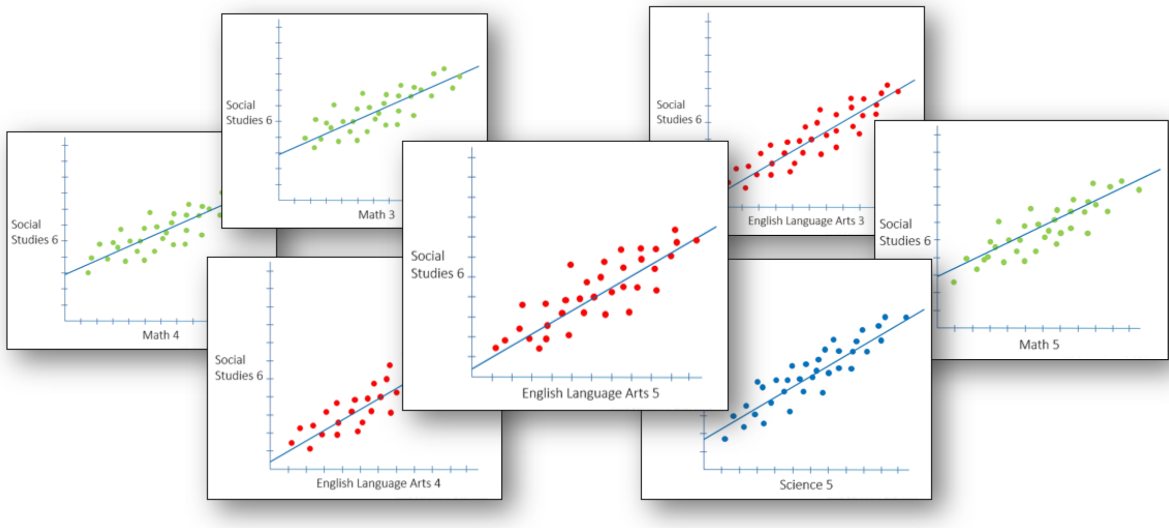
To illustrate how the predictive model uses a student's prior testing history, consider all students who tested in Science in grade 8 in a given year. There isn't a Science test in the immediate prior grade. However, these students might have a number of prior test scores in Math and ELA in grades 3–7, as well as Science in grade 5. These prior test scores have a relationship with Science, meaning that how students performed on these tests can predict how the students perform on Science in grade 8. The growth model does not assume what the predictive relationship will be; instead, the actual relationships observed by the data define the relationships. This is shown in Figure 1 below where each dot represents a student's prior score on Math 7 plotted with their score on Science 8. The best-fit line indicates how students with a certain prior score on Math 7 tend to score, on average, on Science 8. This illustration is based on one prior test; the predictive model uses many prior test scores from different subjects and grades.

Figure 1: Test Scores from One Assessment Have a Predictive Relationship to Test Scores from Another Assessment



Some subjects and grades will have a greater relationship to Science in grade 8 than others; however, the other subjects and grades still have a predictive relationship. For example, prior Math scores might have a stronger predictive relationship to Science scores in grade 8 than it would have to prior ELA scores, but how a student performs on the ELA test typically provides an idea of how the student is expected to perform on average on Science. This is shown in Figure 2 below, where there are a number of different tests that have a predictive relationship with Science in grade 8. All these relationships are considered together in the predictive model with some tests weighted more heavily than others.

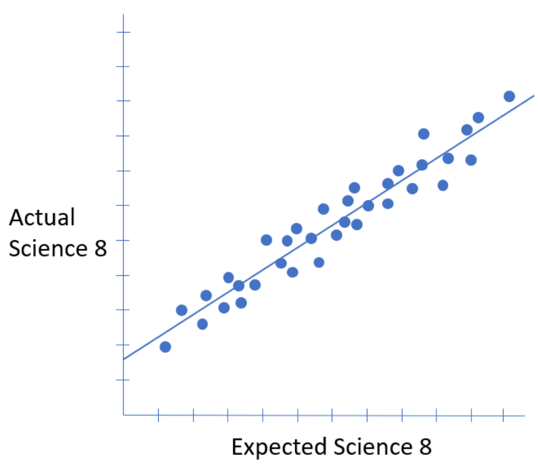
Figure 2: Relationships Observed in the Statewide Data Inform the Predictive Model



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

Each student receives an expected score based on their own prior testing history. In practical terms, the expected score represents the student's entering achievement because it is based on all prior testing information to date. Figure 3 below shows the relationship between expected and actual scores for a group of students.

Figure 3: Relationship Expected Score and Actual Score for Selected Subject and Grade



The expected scores can be aggregated to a specific district or school and then compared to the students' actual scores. In other words, growth is calculated by comparing the difference between the exiting achievement (or average actual score) and the entering achievement (or average expected score) for a group of students. The actual score and expected score are reported in scale score units.

2.2.3 Technical Description of the District, School, and Classroom Models

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

The score to be predicted serves as the response variable (y, the dependent variable).

The covariates (x terms, predictor variables, explanatory variables, independent variables) are scores on tests the student has taken in previous years from the response variable.

There is a categorical variable (class variable, grouping variable) to identify the district, school, or teachers from whom the student received instruction in the subject, grade, and year of the response variable (y).

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

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i \tag{1}$$

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

$$\hat{y}_i = \hat{\mu}_y + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots \tag{2}$$

Two difficulties must be addressed in order to implement the prediction model. First, not all students will have the same set of predictor variables due to missing test scores. Second, because the predictive model is an ANCOVA model, the estimated parameters are pooled within group (district, school, or classroom). 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,

$$C = \begin{bmatrix} c_{yy} & c_{yx} \\ c_{xy} & C_{xx} \end{bmatrix} \tag{3}$$

This matrix is estimated using the EM (expectation maximization) algorithm for estimating covariance matrices in the presence of missing data available in SAS/STAT® (although no imputation is actually used). It should also be noted that, due to this being an ANCOVA model, *C* is a pooled-within group (district or school) covariance matrix. This is accomplished by providing scores to the EM algorithm that are centered around group means (that is, the group means are subtracted from the scores) rather than around grand means. Obtaining *C* is an iterative process since group means are estimated within the EM algorithm to accommodate missing data. Once new group means are obtained, another set of scores is fed into the EM algorithm again until C converges. This overall iterative EM algorithm is what accommodates the two difficulties mentioned above. Only students who had a test score for the response variable in the most recent year and who had at least two predictor variables (for grades 4 and 5) or three predictor variables (for all other assessments) are included in the estimation. Given such a matrix, the vector of estimated regression coefficients for the prediction equation (2) can be obtained as:

$$\hat{\beta} = C_{xx}^{-1} c_{xy} \tag{4}$$

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

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

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

$$\hat{y}_i = \hat{\mu}_y + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots \tag{5}$$

The $\hat{y}_i$ term is nothing more than a composite of all the student's past scores. It is a one-number summary of the student's level of achievement prior to the current year, and this term is called the expected score or entering achievement in the web reporting. The different prior test scores making up this composite are given different weights (by the regression coefficients, the $\hat{\beta}$ terms) in order to maximize its correlation with the response variable. Thus, a different composite would be used when the response variable is Mathematics than when it is ELA, for example. Note that the $\hat{\alpha}_j$ term is not included in the equation. Again, this is because $\hat{y}_i$ represents prior achievement before the effect of the current district, school or classroom.

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

$$y_i = \gamma_0 + \gamma_1 \hat{y}_i + \alpha_j + \epsilon_i \tag{6}$$

In the predictive model, the effects (α_j) are considered random effects. Consequently, the $\hat{\alpha}_j$ terms are obtained by shrinkage estimation (empirical Bayes).[2] The regression coefficients for the ANCOVA model are given by the γ terms.

2.3 Projection Model

2.3.1 Overview

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

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

2.3.2 Technical Description

The statistical model that is used as the basis for the projections is, in traditional terminology, an analysis of covariance (ANCOVA) model. This model is the same statistical model used in the predictive model applied at the school level described in Section 2.2.3. In the projection model, the score to be projected serves as the response variable (y), the covariates (x terms) are scores on tests the student has already taken, and the categorical variable is the school at which the student received instruction in the subject, grade, and year of the response variable (y). Algebraically, the model can be represented as follows for the i^{th} student.

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i \tag{7}$$

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

$$\hat{y}_i = \hat{\mu}_y \pm \hat{\alpha}_j + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots + \epsilon_i \tag{8}$$

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

---

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

Two difficulties must be addressed to implement the projections. First, not all students will have the same set of predictor variables due to missing test scores. Second, because this is an ANCOVA model with a school effect i, the regression coefficients must be "pooled-within-school" regression coefficients. The strategy for dealing with these difficulties is the same as described in Section 2.2.3 using equations (3), (4), and (5) and will not be repeated here.

Typically, the parameter estimates are based on the cohort of students who most recently took the assessment. 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 two available predictor scores for grades 4 and 5 or three available predictor scores for all other assessments. In addition to the projected score itself, the standard error of the projection is calculated ($SE(\hat{y}_i)$). Given a projected score and its standard error, it is possible to calculate the probability that a student will reach some specified benchmark of interest (b). Examples are the probability of scoring at least Proficient on a future grade-level or ISASP assessment. The probability is calculated as the area above the benchmark cutoff score using a normal distribution with its mean equal to the projected score and its standard deviation equal to the standard error of the projected score as described below. Φ represents the standard normal cumulative distribution function.

$$Prob(\hat{y}_i \geq b) = \Phi\left(\frac{\hat{y}_i - b}{SE(\hat{y}_i)}\right) \tag{9}$$

2.4 Outputs from the Models

The outputs of the value-added model are available to Iowa educators with user credentials in the EVAAS web application. For 2022-23 reporting, classroom reports will be available in the web reporting for districts that participated in the Roster Verification pilot in Spring 2023. When a teacher has data for the same assessment in multiple districts in the same year, a separate classroom value-added measure will be created for each district.

2.4.1 Predictive Model

The predictive model provides growth measures for districts, schools, and classrooms in the following content areas:

- ISASP English Language Arts in grades 4–11 (4–8 for classroom reporting)
- ISASP Mathematics in grades 4–11 (4–8 for classroom reporting)
- ISASP Science in grades 5, 8, and 10 (5 and 8 for classroom reporting)

2.4.2 Projection Model

Student projection reports are available to predict how a student might perform on future state assessments. More specifically, most grade-level projections are provided only to a student's next tested grade-level assessments based on that student's most recent tested grade, such as projections to grade 5 for students who most recently tested in grade 4. Science projections are made to the next grade in which a Science assessment is administered. Projections are made to the performance levels Proficient and Advanced, and the individual cut scores depend on each subject and grade. To summarize, the following projections are available for students who meet the reporting criteria:

- ISASP English Language Arts in grades 4–11
- ISASP Mathematics in grades 4–11
- ISASP Science in grades 5, 8, and 10

3 Expected Growth

3.1 Overview

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

A detailed explanation of expected growth and how it is calculated is provided to ensure that appropriate interpretations are made when using the growth measures.

3.2 Technical Description

The predictive model calculates expected growth based on the empirical results of student testing data. The predictive model does not assume a particular amount of growth or assign expected growth in advance of the assessment being taken by students. The model defines expected growth within a year. This means that expected growth is always relative to how students' achievement has changed in the most recent year of testing rather than a fixed year in the past.

More specifically, **in the predictive model, expected growth means that students within a district, school, or classroom made the same amount of growth as students within the average district, school, or classroom in the state for that same year, subject, and grade**.

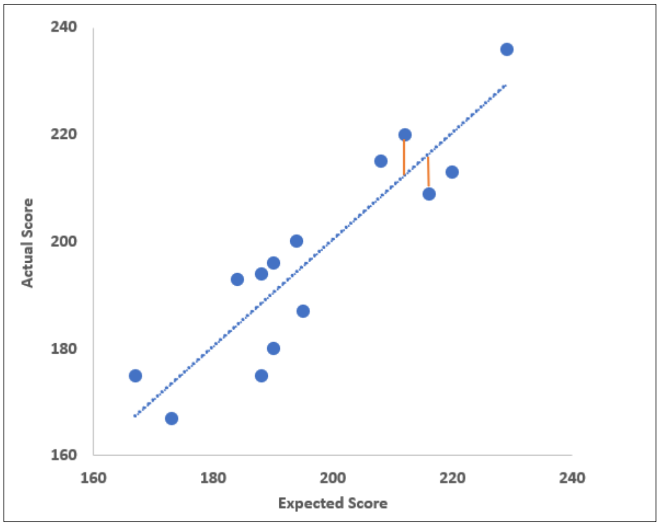
Growth measures tend to be centered on expected growth every year with approximately half of the district/school/classroom estimates above zero and approximately half of the district/school/classroom estimates below zero.

A change in assessments or scales from one year to the next does not present challenges to calculating expected growth. The predictive model already uses prior test scores from different scales to calculate the expected score. When assessments change over time, expected growth is still based on the relative change in achievement from one point in time to another.

3.3 Illustrated Example

In the predictive model, expected growth uses actual results from the most recent year of assessment data and considers the relationships from the most recent year with prior assessment results. Figure 4 below provides a simplified example of how growth is calculated in the predictive model. The graph plots each student's actual score with their expected score. Each dot represents a student, and a best-fit line shows the difference between all students' actual and expected scores. Collectively, the best-fit line indicates what expected growth is for each student – given the student's expected score, expected growth is met if the student scores the corresponding point on the best-fit line. Conceptually, with the best-fit line minimizing the difference between all students' actual and expected scores, the growth expectation is defined by the average experience. Note that the actual calculations differ slightly since this is an ANCOVA model where the students are expected to see the average growth as seen by the experience with the average group (district, school, or classroom).

Figure 4: Intra-Year Approach Example for the Predictive Model



4 Classifying Growth into Categories

4.1 Overview

It can be helpful to classify growth into different levels for interpretation and context, particularly when the levels have statistical meaning. Iowa's results include five categories for districts, schools, and classrooms. These categories are defined by a range of values related to the growth measure and its standard error, and they are known as growth indicators in the web application.

4.2 Use Standard Errors Derived from the Models

As described in the modeling approaches section, the growth model provides an estimate of growth for a district, school, or classroom in a particular subject, grade, and year as well as that estimate's standard error. The standard error is a measure of the quantity and quality of student data included in the estimate, such as the number of students and the occurrence of missing data for those students. Standard error is a common statistical metric reported in many analyses and research studies because it yields important information for interpreting an estimate that is, in this case, the growth measure relative to expected growth. Because measurement error is inherent in any growth or value-added model, the standard error is a critical part of the reporting. **Taken together, the growth measure and standard error provide educators and policymakers with critical information about the certainty that students in a district, school, or classroom are making decidedly more or less than the expected growth.** Taking the standard error into account is particularly important for reducing the risk of misclassification (for example, identifying a school as ineffective when it is truly effective).

The standard error also takes into account that districts, schools, and classrooms with the same number of students might have students with very different amounts of prior testing history. Due to this variation, the standard errors in a given subject, grade, and year could differ significantly among districts, schools, and classrooms. Standard errors are important because of this variability in available data and act as an important protection for districts, schools, and classrooms in the growth reporting.

4.3 Define Growth Indicators in Terms of Standard Errors

Common statistical usage of standard errors indicates the precision of an estimate and whether that estimate is statistically significantly different from an expected value. The growth reports use the standard error of each growth measure to determine the statistical evidence that the growth measure is different from expected growth. For Iowa's EVAAS growth reporting, this is essentially when the growth measure is more than or less than two standard errors above or below expected growth or, in other words, when the growth index is more than +2 or less than -2. These definitions then map to growth indicators in the reports themselves, such that there is statistical meaning in these categories. The categories and definitions are illustrated in the following section.

4.4 Illustrated Examples of Categories

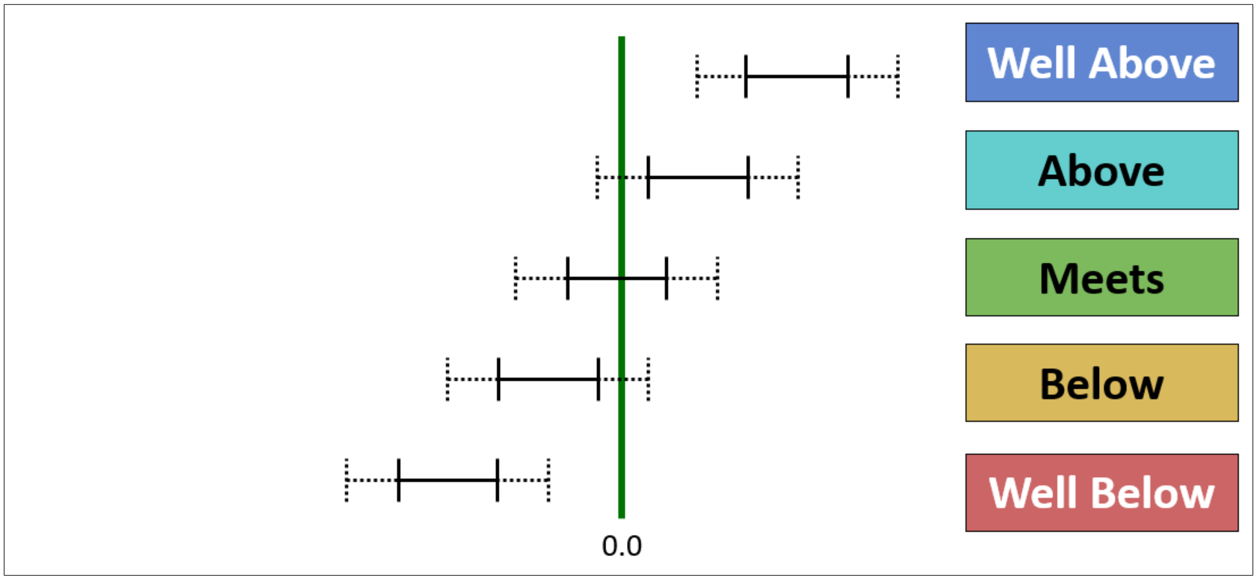
There are two ways to visualize how the growth measure and standard error relate to expected growth and how these can be used to create categories.

The first way is to frame the growth measure relative to its standard error and expected growth at the same time. For district, school, and classroom reporting, the categories are defined as follows:

- **Well Above** indicates that the growth measure is two standard errors or more above expected growth (0). This level of certainty represents significant evidence that students made more growth than expected.
- **Above** indicates that the growth measure is at least one but less than two standard errors above expected growth (0). This is moderate evidence that students made more growth than expected.
- **Meets** indicates that the growth measure is less than one standard error above expected growth (0) but no more than one standard error below expected growth (0). This is evidence that students made growth as expected.
- **Below** indicates that the growth measure is more than one but no more than two standard errors below expected growth (0). This is moderate evidence that students made less growth than expected.
- **Well Below** is an indication that the growth measure is two standard errors or more below expected growth (0). This level of certainty represents significant evidence that students made less growth than expected.

Figure 5 below shows visual examples of each category. The green line represents the expected growth. The black line represents the range of values included in the growth measure plus and minus *one* standard error. The dotted black line extends the range of values to the growth measure plus and minus *two* standard errors. If the dotted black line is completely above expected growth, then there is significant evidence that students made more than expected growth, which represents the Well Above category. Conversely, if the dotted black line is completely below expected growth, then there is significant evidence that students made less than expected growth, which represents the Well Below category. The Above and Below categories indicate, respectively, that there is moderate evidence that students made more than expected growth and less than expected growth. In these categories, the black line is completely above or below expected growth but not the dotted black line. In addition, Meets indicates that there is evidence that students made growth as expected as both the black and dotted lines cross the line indicating expected growth.

Figure 5: Visualization of Growth Categories with Expected Growth, Growth Measures, and Standard Errors



This graphic is helpful in understanding how the growth measure relates to expected growth and whether the growth measure represents a statistically significant difference from expected growth.

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

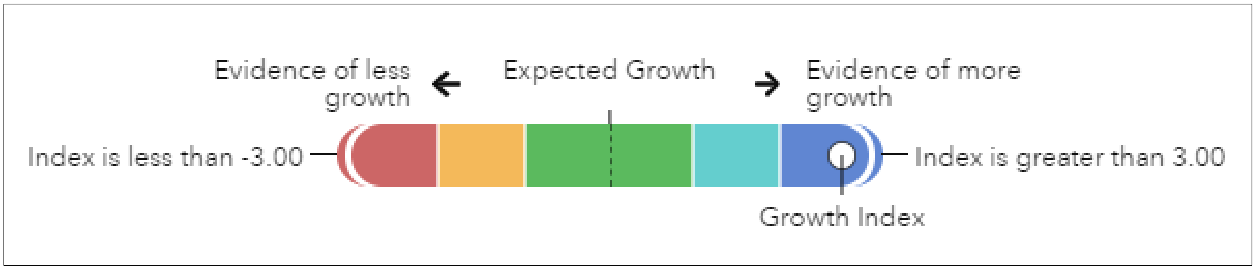
$$Growth\ Index = \frac{Growth\ Measure - Expected\ Growth}{Standard\ Error\ of\ the\ Growth\ Measure} \tag{10}$$

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

- Well Above indicates significant evidence that students made more growth than expected. The growth index is 2 or greater.
- Above indicates moderate evidence that students made more growth than expected. The growth index is between 1 and 2.
- Meets indicates evidence that students made growth as expected. The growth index is between -1 and 1.
- Below indicates moderate evidence that students made less growth than expected. The growth index is between -2 and -1.
- Well Below indicates significant evidence that students made less growth than expected. The growth index is less than -2.

This is represented in the growth indicator bar in Figure 6, which is similar to what is provided in the District and School Value-Added Reports in the EVAAS web application. The black dotted line represents expected growth. The color-coding within the bar indicates the range of values for the growth index within each category.

Figure 6: Sample Growth Indicator Bar



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

4.5 Rounding and Truncating Rules

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

5 Input Data Used in EVAAS Growth Model

5.1 Assessment Data

For the analysis and reporting based on the 2022-23 school year, EVAAS received assessment data for use in the growth and/or projection models. Iowa Assessments data were received for the 2014-15 through 2017-18 school years, and ISASP data were received for the 2018-19 through 2022-23 school years (with the exception of 2019-20 due to the COVID-19 pandemic):

- ISASP English Language Arts Grades 3–11
- ISASP Mathematics Grades 3–11
- ISASP Science Grades 5, 8, and 10
- Iowa Assessments Reading Grades 3–11
- Iowa Assessments Mathematics Grades 3–11
- Iowa Assessments Science Grades 3–11

Assessment files provide the following data for each student score:

- Student Identifiers (StateStudentID, Student First Name, Student Last Name, Student DOB)
- Test Taken
- Tested Subject
- Tested Grade
- Tested Period (Fall/Midyear/Spring – Iowa Assessments only)
- District Number
- School Number
- Scale Score[3]
- Performance Level (ISASP Only)

5.2 Student Information

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

- Gender (Male, Female, Non-Binary, Unknown)
- Race/Ethnicity (Asian, Black or African American, Hispanic/Latino, Two or More Races, American Indian/Alaska Native, Native Hawaiian or Other Pacific Islander, White, Unknown)
- English Learner (Yes, No, Unknown)

[3] Spring converted scores were received and used for Iowa Assessments through the 2017-18 school year

- Foster Care (Yes, No, Unknown)
- Free or Reduced-Price Lunch (Yes, No, Unknown)
- Gifted (Yes, No, Unknown)
- Homeless (Yes, No, Unknown)
- Military Connected (Yes, No, Unknown)
- Section 504 (Yes, No, Unknown)
- Students with Disabilities (Yes, No, Unknown)
- Partial Academic Year (Yes, No, Unknown)
- Chronically Absent (Yes, No, Unknown)

5.3 Student-Teacher Linkages

The Iowa Department of Education provided SAS with **student, teacher, and course data from Fall and Winter collections** that could be used to prepopulate the Roster Verification application. The Roster Verification application allowed educators and administrators to verify which students should be connected to which teachers for use in classroom reporting. This data included the following categories:

- Teacher-Level Identification
- Teacher Name
- Teacher Folder Numbers
- Student Information
- Student Last Name
- Student First Name
- StateStudentID
- Grade
- Course Titles to identify sections associated with instruction for ELA, Mathematics, and Science tested subjects in grades 4–8
- District and School Information (Numbers)
- Percentage of instructional responsibility (derived from dividing 100% across the number of teachers providing instruction to a student in a tested subject and grade)

This data was provided for Mathematics and ELA in grades 4–8 and Science in grades 5 and 8 to align for the assessments included in classroom reporting.

6 Business Rules

6.1 Assessment Verification for Use in Growth Models

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

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

These criteria are checked annually for each assessment prior to use in any growth model. These criteria are explained in more detail below.

6.1.1 Stretch

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

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

6.1.2 Relevance

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

6.1.3 Reliability

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

6.2 Pre-Analytic Processing

6.2.1 Missing Grade

In Iowa, the grade used in the analyses and reporting is the tested grade, not the enrolled grade. If a grade is missing on a grade-level test record (meaning 3–11), then that record will be excluded from all analyses. The grade is required to include a student's score in the appropriate part of the models.

6.2.2 Use of Spring Converted Score for Iowa Assessments

The spring converted scale score was used for the Iowa Assessments administered through the 2017-18 school year.

6.2.3 Off-Grade Science Test Takers

All student records for ISASP Science with grade levels other than 5, 8, or 10 were removed from the analysis. Iowa Assessments Science scores through the 2017-18 school year span all grade levels and were retained.

6.2.4 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. The record with the most demographic information is prioritized in selecting the record to keep.

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

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

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

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

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

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

6.2.8 Students with Records That Have Unexpected Grade Level Changes

If a student skips more than one grade level (for example, a student moves from sixth in 2021 to ninth in 2022) or is moved back by one grade or more (for example, a student moves from fourth in 2021 to third in 2022) 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.

6.2.9 Outliers

Student assessment scores are checked each year to determine whether they are outliers in context with all the other scores in a reference group of scores from the individual student. These reference scores are weighted differently depending on proximity in time to the score in question. Scores are checked for outliers using related subjects as the reference group. For example, when searching for outliers for Math test scores, all Math assessments are examined simultaneously, and any scores that appear inconsistent, given the other scores for the student, are flagged.

Scores are flagged in a conservative way to avoid excluding any student scores that should not be excluded. Scores can be flagged as either high or low outliers. Once an outlier is discovered, then that outlier will not be used in the analysis, but it will be displayed in the student testing history on the EVAAS web application.

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

Considerations included in outlier detection are:

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

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

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

For low-end outliers, the rules are:

- The percentile of the score must be below 50.
- The t-value must be below -3.5 for Math and ELA 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 Science 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 range from 10 to 90 with the ranges of the individual percentile score.

For high-end outliers, the rules are:

- The percentile of the score must be above 50.
- The t-value must be above 4.5 for Math and ELA 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 Science 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 scores in the comparison score average.

6.2.10 Linking Records Over Time

Each year, EVAAS receives data files that include student assessment data and file formats. These data are checked each year prior to incorporation into a longitudinal database that links students over time. Student test data and demographic data are checked for consistency from year to year to ensure that the appropriate data are assigned to each student. Student records are matched over time using all data provided to EVAAS.

6.3 Growth Models

6.3.1 Students Included in the Analysis

As described in Pre-Analytic Processing (Section 6.2), student scores might be excluded due to the business rules, such as outlier scores.

For the predictive and projection models, students are with a school_id value of 0 and/or a district_id value of 0 in the year of reporting are excluded in accordance with a policy decision made by the Iowa Department of Education.

For the predictive and projection models, a student must have the required number of predictor scores – two for grades 4 and 5 and three for all other assessments - that can be used in the analysis, all of which cannot be deemed outliers. (See Section 6.2.9 on Outliers.) These scores can be from any year, subject, and grade that are used in the analysis. In other words, the student's expected score can incorporate other subjects beyond the subject of the assessment being used to measure growth. The required 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 minimum, then that student is excluded from the analyses. It is important to note that not all students have to have the same prior test scores; they need only to have some subset that were used in the analysis. Since the predictive model does not determine growth based on consecutive grade movement on tests, students do not need to stay in one cohort from one year to the next. That said, if a student is retained and retakes the same test, then that prior score on the same test will not be used as a predictor for the same test as a response in the predictive model. This is mainly because very few students used in the models have a prior score on the same test that could be used as a predictor. In fact, in the predictive model, it is typically the case that a prior test is only considered to be a possible predictor when at least 50% of the students used in that model also have those prior test scores.

The predictive and projection models exclude first-time English Learner test takers who have no prior testing history. These students are included in future years if they have prior scores that can be used in the analysis.

6.3.2 Minimum Number of Students to Receive a Report

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

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

Classroom reporting requires at least 5 students and five full-time effective (FTE) students to produce a growth measure. FTE counts consider both the number of students and the percentage of instructional responsibility for the teacher. For example, five students with 100% instructional responsibility results in 5 FTEs, and 10 students with 50% instructional responsibility results in 5 FTEs.