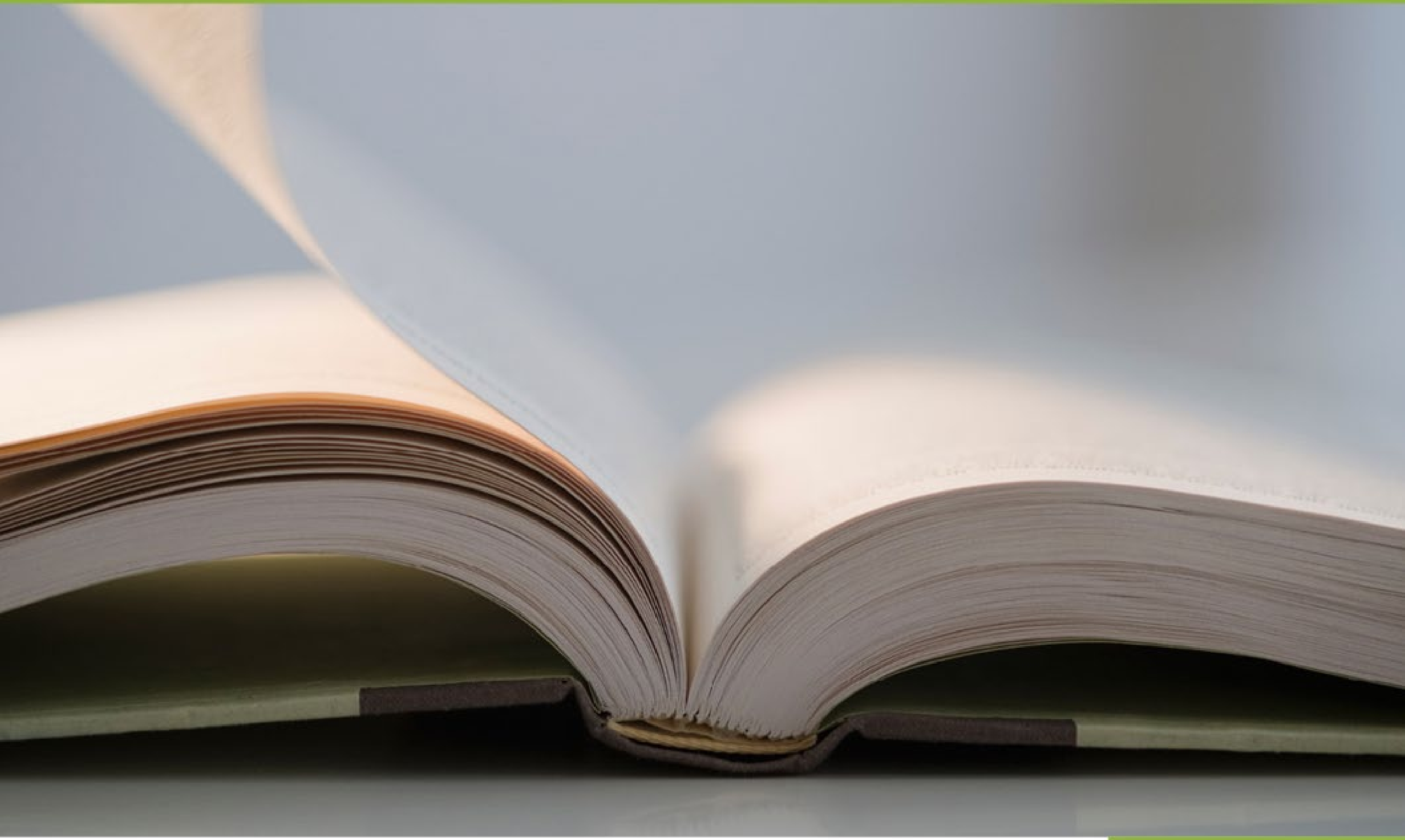
SAS® EVAAS

Statistical Models and Business Rules

Prepared for the Idaho State Board of Education



THE POWER TO KNOW.

Contents

1 Introduction to Idaho's Value-Added Reporting

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

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

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

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

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

- Conceptual and technical explanations of analytic models
- Definition of expected relative progress to support the calculation of relative progress results
- Classifying relative progress into categories for interpretation of results
- Input data
- Business rules

These reports are delivered through the EVAAS web application and the Idaho K-12 Education Data Dashboard. Although the underlying statistical models and business rules supporting these reports are sophisticated and comprehensive, the web reports are designed to be user-friendly and visual so that educators and administrators can quickly identify strengths and opportunities for improvement and then use these insights to inform curricular, instructional, and planning supports.

2 Statistical Models

2.1 Overview of Statistical Models

The conceptual explanation of value-added reporting is simple: compare students' exiting achievement with their entering achievement over two points in time. In practice, however, measuring student relative progress 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 Idaho, can account for these complexities and scenarios.

Idaho's value-added models and the resulting relative progress measures 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 relative progress measures. Furthermore, because certain students are more likely to have missing test scores than others, this approach provides less biased relative progress measures than models that cannot accommodate student 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 relative progress measures, particularly those that are not tested every year.

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

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

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

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

Based on Idaho's state assessment program, there are two approaches to providing district and school relative progress measures. Both approaches are conceptually similar in that they measure relative progress.

- **The gain model (also known as the multivariate response model or MRM)** is sometimes used when assessments are administered in the same content area in consecutive grades. For the 2020-21 reporting, this model is used for the Idaho Reading Indicator (IRI) in grades K-3 and for Idaho Standards Achievement Test (ISAT) assessments for Math and English Language Arts (ELA) in grades 5–8. For the 2021-22 reporting, this model is only used for IRI assessments in grades K-3.
- **The predictive model (also known as univariate response model or URM)** can be used when a test is given in non-consecutive grades or when performance from previous tests is used to predict performance on another test. Like the gain model, this approach can also be used in cases where assessments are administered in the same content area in consecutive grades. For the 2020-21 reporting, this model is used for ISAT Math and ELA in grades 3 and 4 (because only prior IRI scores are available for those students), ISAT Math and ELA in grade 10, PSAT NMSQT in grade 10, and SAT in grade 11. Beginning with the 2021-22 reporting, this approach will be used for all assessments other than IRI.

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

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

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

2.2 Gain Model

2.2.1 Overview

The gain model measures relative progress between two points in time for a group of students. **More specifically, the gain model measures the change in relative achievement for a group of students based on the statewide achievement from one point in time to the next.** For state summative assessments, relative progress is typically measured from one year to the next using the available consecutive grade assessments. For IRI assessments, relative progress is measured from the fall administration of IRI to the spring administration of IRI within the same grade. Expected relative progress means that students maintained their relative achievement among the population of test-takers, and more details are available in Section 3.

There are two separate analyses for EVAAS reporting based on the gain model: one each for districts and schools. The district and school models are essentially the same; they perform well with the large numbers of students characteristic of districts and most schools.

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

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

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

2.2.2 Why the Gain Model is Needed

A common question is how this approach differs from measuring the changes between the current year's scores and prior year's scores for a group of students. The example in Figure 1 illustrates these differences.

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

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

Student	Previous Score	Current Score	Gain
1	51.9	74.8	22.9
2	37.9	46.5	8.6
3	55.9	61.3	5.4
4	52.7	47.0	-5.7
5	53.6	50.4	-3.2
6	23.0	35.9	12.9
7	78.6	77.8	-0.8
8	61.2	64.7	3.5
9	47.3	40.6	-6.7
10	37.8	58.9	21.1
Column Mean	**50.0**	**55.8**	**5.8**
Difference between Current and Previous Score Means			**5.8**

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

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

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

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

2.2.3 Common Scale in the Gain Model

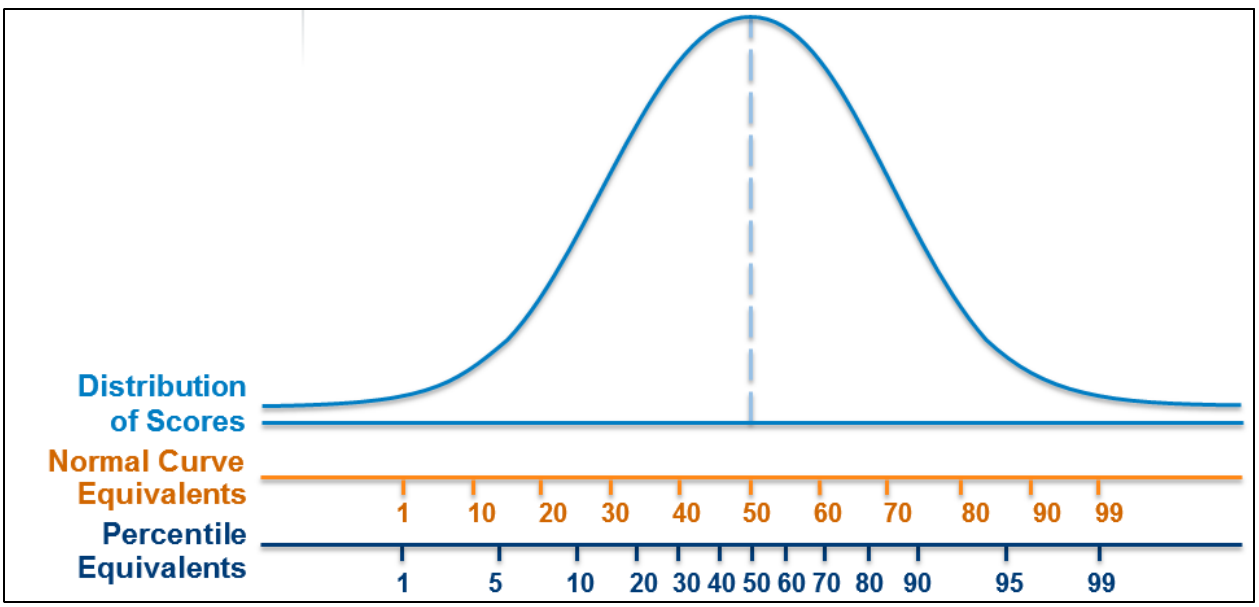
2.2.3.1 Why the Model Uses Normal Curve Equivalents

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

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

However, NCEs have a critical advantage over percentiles for measuring relative progress: NCEs are on an equal-interval scale. This means that for NCEs, unlike percentile ranks, the distance between 50 and 60 is the same as the distance between 80 and 90. This difference between the distributions is evident below in Figure 2.

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



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

Each NCE distribution is based on a specific assessment, test, subject, and time point. For example, the NCE distribution for 2021 Math in grade 5 is constructed separately from the NCE distribution for 2021 Math in grade 6.

2.2.3.2 How to Calculate NCEs in the Gain Model

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

Table 1 provides an example of how the gain model converts scale scores to NCEs. In a given subject, grade, and year, the tabulation shows, for each given score, the percentage of students who scored that score ("Percent"). The table also tabulates the "Cumulative Frequency as the number of students who made that score or lower and its associated percentage ("Cumulative Percent").

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

Table 1: Converting Tabulated Test Scores to NCE Values

Score	Percent	Cumulative Percent	Percentile Rank	Z-Score	NCE
418	3.1	36.9	35.4	-0.375	42.10
420	3.3	40.2	38.5	-0.291	43.87
423	3.3	43.5	41.8	-0.206	45.66
425	3.4	46.9	45.2	-0.121	47.46
428	3.5	50.4	48.6	-0.035	49.27
430	3.5	53.9	52.1	0.053	51.12
432	3.6	57.4	55.7	0.143	53.00

NCEs are obtained from the percentile ranks using the normal distribution. The table of the standard normal distribution (found in many textbooks[2]) or computer software (for example, a spreadsheet) provides the associated Z-score from a standard normal distribution for any given percentile rank. NCEs are Z-scores that have been rescaled to have a "percentile-like" scale. As mentioned above, the NCE

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

distribution is scaled so that NCEs exactly match the percentile ranks at 1, 50, and 99. To do this, each Z-score is multiplied by approximately 21.063 (the standard deviation on the NCE scale) and then 50 (the mean on the NCE scale) is added.

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

2.2.4 Technical Description of the Gain Model

2.2.4.1 Definition of the Linear Mixed Model

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2.4.2 District and School Models

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

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

Algebraically, the district gain model is represented as:

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

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

The school gain model is represented as:

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

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

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

Solving the mixed model equations for the district or school gain model produces a vector b that contains the estimated mean score for each school (in the school model), subject, grade, and year. To obtain a value-added measure of average student relative progress, a series of computations can be done using the students from a school in a particular year and their prior and current testing data. The model produces means in each subject, grade, and year that can be used to calculate differences in order to obtain gains. Because students might change schools from one year to the next (in particular when transitioning from elementary to middle school, for example), the estimated mean score for the prior year/grade uses students who existed in the current year of that school. Therefore, mobility is taken into account within the model. Relative progress 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 relative progress 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 relative progress measures are reported along with standard errors, and these can be obtained by taking the square root of equation (6) as described above.

2.2.4.3 Accommodations to the Gain Model for Missing 2019-20 Data Due to the Pandemic

2.2.4.3.1 Overview

In spring 2020, the COVID-19 pandemic disrupted instruction and caused the cancellation of statewide summative assessments for the 2019-20 school year. As a result, scores are not available for Idaho's ISAT assessments based on the 2019-20 school year, and it is not possible to measure relative progress on ISAT from the 2018-19 to the 2019-20 school years or from the 2019-20 to the 2020-21 school years. For the gain model based on ISAT assessments, the 2020-21 reporting measures relative progress from the 2018-19 school year to the 2020-21 school year.

From a technical perspective, the district and school gain model for ISAT assessments is essentially the same as in a more typical year except that relative progress is measured over two years rather than one year. Because EVAAS measures the change in *relative* achievement based on the statewide population of test-takers, the measures represent progress *relative* to the average relative progress observed in the state. In other words, a drop in achievement or proficiency rates due to lost instructional time does not correspond to a drop in relative progress. District and school relative progress measures are still relative to the state average, and expected relative progress is based on students' maintaining their achievement among the population of test-takers.

That said, the interpretation of these relative progress measures changes slightly. Because the models provide two-year relative progress measures, the relative progress measure for grades where students transition from one school to another will then include relative progress from the feeder school(s) as well as the receiver school. For example, a middle school with grades 6–8 could receive a relative progress measure for sixth grade based on the students' relative progress in sixth grade as well as their relative progress from the feeder elementary school(s) in fifth grade.

In other words, it is not possible to parse out the individual contribution of the middle school in sixth grade apart from those from the elementary school(s) in fifth grade because of the missing year of test scores. For the district-level relative progress measures and for the non-transition grades, the two-year relative progress measures are still solely representative of relative progress within the specific district and the non-transition grades for the school are still solely representative of relative progress within the specific school.

Despite these differences, the conceptual explanation of the 2020-21 relative progress measures remains the same: these relative progress measures compare students' exiting achievement with their entering achievement over two points in time and provide a measure of relative progress.

Because IRI assessments were administered in both fall 2020 and spring 2021, the results for these assessments can be interpreted as they are in a more typical school year, with relative progress measured between those assessment administrations.

For the 2021-22 analysis, the gain model returned to more typical circumstances for IRI with the availability of prior scores from 2020-21.

2.3 Predictive Model

2.3.1 Overview

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

Like the gain model, there are two separate analyses for EVAAS reporting based on the predictive model: one each for districts and schools. The district and school models are essentially the same.

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 relative progress, regression models identify the relationship between prior test performance and actual test performance for a given course. In more technical terms, the predictive model is known as the univariate response model (URM), a linear mixed model and, more specifically, an analysis of covariance (ANCOVA) model.

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

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

2.3.2 Conceptual Explanation

As mentioned above, the predictive model is ideal for assessments given in non-consecutive grades and in cases where the available prior assessment data is from a different type of assessment. Consider all students who tested in ISAT Math in grade 10 in a given year. The gain model is not possible since there isn't a Math test in the immediate prior grade. However, these students might have a number of prior test scores spanning multiple subject areas from prior years. These prior test scores have a relationship with ISAT Math in grade 10, meaning that how students performed on these tests can predict how the students perform on that assessment. The relative progress model does not assume what the predictive relationship will be; instead, the actual relationships observed by the data define the relationships.

Some subjects and grades will have a greater relationship to a given assessment 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 ISAT Math in grade 10 than prior Math scores, but how a student reads and performs on prior ELA tests typically provides an idea of how we might expect a student to perform on average on future Math tests. All of these relationships are considered together in the predictive model with some tests weighted more heavily than others.

Note that the prior test scores do not need to be on the same scale as the assessment being measured for student relative progress. Just as height (reported in inches) and weight (reported in pounds) can predict a child's age (reported in years), the 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.

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

2.3.3 Technical Description of the District and School Models

The predictive model has similar approaches for districts and schools. 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 or school 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.

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i \tag{13}$$

Two difficulties must be addressed in order to implement the predictive 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 or school). 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,

$$C = \begin{bmatrix} c_{yy} & c_{yx} \\ c_{xy} & C_{xx} \end{bmatrix} \tag{14}$$

Note that *C* in equation (14) is not the same as *C* in equation (4). This matrix is estimated using the EM (expectation maximization) algorithm for estimating covariance matrices in the presence of missing data available in SAS/STAT® (although no imputation is actually used). It should also be noted that, due to this being an ANCOVA model, *C* is a pooled-within group (district, school, or teacher) covariance matrix. This is accomplished by providing scores to the EM algorithm that are centered around group means (i.e., the group means are subtracted from the scores) rather than around grand means. Obtaining *C* is an iterative process since group means are estimated within the EM algorithm to accommodate missing data. Once new group means are obtained, another set of scores is fed into the EM algorithm again until *C* converges. This overall iterative EM algorithm is what accommodates the two difficulties mentioned above. Only students who had a test score for the response variable in the most recent year and who had at least three predictor variables are included in the estimation. Given such a matrix, the vector of estimated regression coefficients for the projection equation (15) can be obtained as:

$$\hat{\beta} = C_{xx}^{-1} c_{xy} \tag{15}$$

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

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

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

$$\hat{y}_i = \hat{\mu}_y + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots \tag{16}$$

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

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

$$y_i = \gamma_0 + \gamma_1\hat{y}_i + \alpha_j + \epsilon_i \tag{17}$$

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

2.3.3.1 Accommodations to the Predictive Model for Missing 2019-20 Data due to the Pandemic

In spring 2020, the COVID-19 pandemic disrupted instruction and led to the cancellation of spring 2020 assessments. As a result, it is not possible to measure relative progress from the 2018-19 to the 2019-20 school years. For the predictive model, the 2020-21 reporting measures relative progress using students' predictors through the 2018-19 school year where available and then compares to their performance on the 2020-21 assessment. In the 2021-22 analysis the available prior scores through 2020-21 are used, again with the exception of the unavailability of 2019-20 scores.

As a reminder, the predictive model is sometimes used to measure relative progress for assessments given in non-consecutive grades, such as ISAT Math and ELA in grade 10. Because these assessments are not administered every year, *it has always been possible* that students do not have any test scores in the *immediate* prior year. The model can provide a robust estimate of students' entering achievement for the course by using all other available test scores from other subjects, grades, and years.

In other words, the predictive model does not require any technical adaptations to account for the missing year of data and the interpretation of the results is similar to a typical year of reporting.

---

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

2.4 Projection Model

2.4.1 Overview

The longitudinal data sets used to calculate relative progress 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 Level 3 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 available for state assessments as well as to college readiness assessments.

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

2.4.2 Technical Description

The statistical model that is used as the basis for the projections is, in traditional terminology, an analysis of covariance (ANCOVA) model. This model is the same statistical model used in the predictive model applied at the school level described in Section 2.3.3. In the projection model, the score to be projected serves as the response variable (y), the covariates (x terms) are scores on tests the student has already taken, and the categorical variable is the school at which the student received instruction in the subject, grade, and year of the response variable (y). Algebraically, the model can be represented as follows for the i^{th} student.

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i \tag{18}$$

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

$$\hat{y}_i = \hat{\mu}_y \pm \hat{\alpha}_j + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots \tag{19}$$

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

Two difficulties must be addressed to implement the projections. First, not all students will have the same set of predictor variables due to missing test scores. Second, because this is an ANCOVA model with a school effect i, the regression coefficients must be "pooled-within-school" regression coefficients. The strategy for dealing with these difficulties is the same as described in Section 2.3.3 using equations (14), (15), and (16) 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 typically made only for students who have at least three available predictor scores. In addition to the projected score itself, the standard error of the projection is calculated ($SE\,(\hat{y}_i)$). Given a projected score and its standard error, it is possible to calculate the probability that a student will reach some specified benchmark of interest (b). Examples are the probability of scoring at least Proficient on a future end-of-grade test or the probability of scoring at least an established college readiness benchmark. The probability is calculated as the area above the benchmark cutoff score using a normal distribution with its mean equal to the projected score and its standard deviation equal to the standard error of the projected score as described below. Φ represents the standard normal cumulative distribution function.

$$Prob(\hat{y}_i \geq b) = \ \Phi\left(\frac{\hat{y}_i - b}{SE\,(\hat{y}_i)}\right) \tag{20}$$

2.5 Outputs from the Models

2.5.1 Gain Model

The gain model is sometimes used for courses where students test in consecutive grade-given tests or for tests administered in the fall and spring. As such, **the gain model provides district and school relative progress measures in the following content areas:**

- IRI in grades K-3 (2020-21 and 2021-22)
- ISAT Math in grades 5–8 (2020-21 only, as the predictive model was used in 2021-22)
- ISAT ELA in grades 5–8 (2020-21 only, as the predictive model was used in 2021-22)

In addition to the mean scores and mean gain for an individual subject, grade, and year, the gain model can also provide cumulative gains across grades for each subject and year. In general, these are all different forms of linear combinations of the fixed effects (and random effects for the teacher model), and their estimates and standard errors are computed in the same manner described above in equations (5) and (6) for district and school models.

2.5.2 Predictive Model

The predictive model can be used in a variety of contexts, including where students test in non-consecutive grade-given tests. As such, for 2020-21 reporting **the predictive model provides relative progress measures for districts and schools in the following content areas:**

- ISAT Math in grades 3, 4, and 10
- ISAT ELA in grades 3, 4, and 10
- PSAT NMSQT and SAT for Evidence-Based Reading and Writing and Mathematics

For the 2021-22 reporting, the predictive model was used for all assessments other than IRI.

2.5.3 Projection Model

Projections are provided to future state assessments as well as college readiness assessments. More specifically, for the 2021-22 reporting, projections are provided for ISAT assessments in grades 4–8 and 10, PSAT NMSQT, and SAT.

3 Expected Relative Progress

3.1 Overview

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

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

3.2 Technical Description

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

More specifically, **in the gain model, expected relative progress means that students maintained the same relative position with respect to the statewide student achievement that year. In the predictive model, expected relative progress means that students with a district, school, or teacher made the same amount of relative progress as students with the average district or school in the state for that same year, subject, and grade**.

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

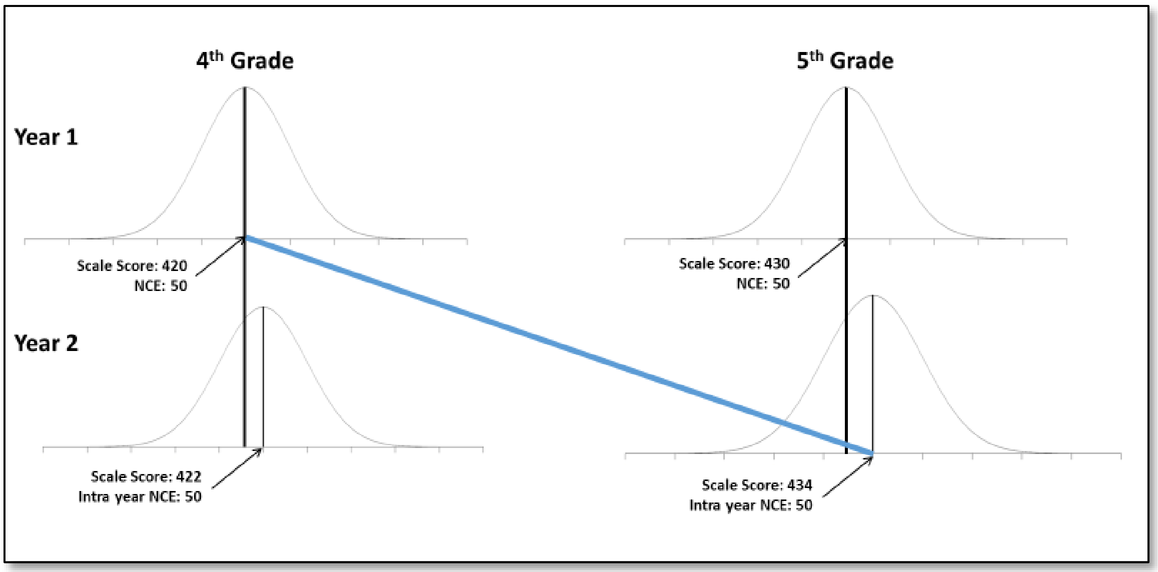
A change in assessments or scales from one year to the next does not present challenges to calculating expected relative progress. Through the use of NCEs, the gain model converts any scale to a relative position, and the predictive model already uses prior test scores from different scales to calculate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

3.3 Illustrated Example

Figure 3 below provides a simplified example of how relative progress is calculated in the gain model when the state achievement increases. The figure has four graphs, each of which plot the NCE distribution of scale scores for a given year and grade. In this example, the figure shows how the gain is calculated for a group of grade 4 students in Year 1 as they become grade 5 students in Year 2. In Year 1, our grade 4 students score, on average, 420 scale score points on the test, which corresponds to the 50th NCE (similar to the 50th percentile). In Year 2, the students score, on average, 434 scale score points on the test, which corresponds to a 50th NCE *based on the grade 5 distribution of scores in Year 2*. The grade 5 distribution of scale scores in Year 2 was higher than the grade 5 distribution of scale scores in Year 1, which is why the lower right graph is shifted slightly to the right. The blue line shows what is required for students to make expected relative progress, which would be to maintain their position at the 50th NCE for grade 4 in Year 1 as they become grade 5 students in Year 2. The relative progress measure for these students is Year 2 NCE – Year 1 NCE, which would be 50 – 50 = 0. Similarly, if a group of students started at the 35th NCE, the expectation is that they would maintain that 35th NCE.

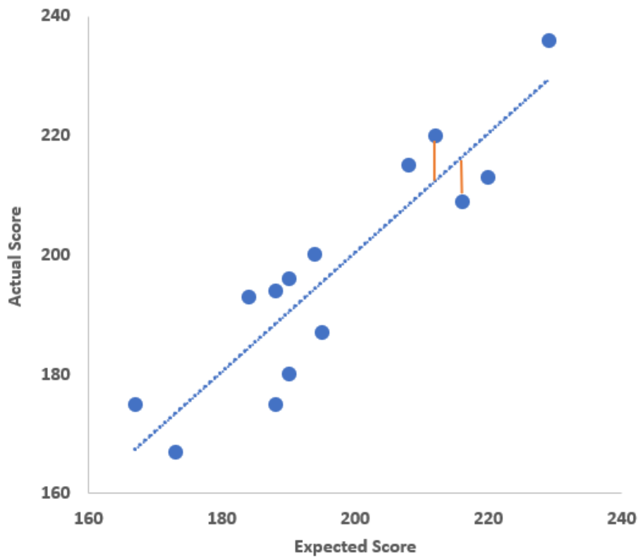
Note that the actual gain calculations are much more robust than what is presented here; as described in the previous section, the models can address students with missing data, team teaching, and all available testing history.

Figure 3: Intra-Year Approach Example for the Gain Model



In contrast, in the predictive model, expected relative progress 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 relative progress is calculated in the predictive model. The graph plots each student's actual score with their expected score. Each dot represents a student, and a best-fit line will minimize the difference between all students' actual and expected scores. Collectively, the best-fit line indicates what expected relative progress is for each student – given the student's expected score, expected relative progress 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 relative progress 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 relative progress as seen by the experience with the average group (district or school).

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



4 Classifying Relative Progress into Categories

4.1 Overview

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

4.2 Use Standard Errors Derived from the Models

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

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

4.3 Define Relative Progress 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 relative progress reports use the standard error of each relative progress measure to determine the statistical evidence that the relative progress measure is different from expected relative progress. For EVAAS reporting, this is essentially when the relative progress measure is more than or less than two standard errors above or below expected relative progress or, in other words, when the relative progress index is more than +2 or less than -2. These definitions then map to relative progress 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 relative progress measure and standard error relate to expected relative progress and how these can be used to create categories.

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

- **Well Above** indicates that the relative progress measure is two standard errors or more above expected relative progress (0). This level of certainty is significant evidence of exceeding the standard for relative progress.
- **Above** indicates that the relative progress measure is one standard error or more above expected relative progress (0). This level of certainty is moderate evidence of exceeding the standard for relative progress.
- **Meets** indicates that the relative progress measure is less than one standard error above expected relative progress (0) and no more than one standard error below it (0). This is evidence of meeting the standard for relative progress.
- **Below** is an indication that the relative progress measure is more than one standard error below expected relative progress (0). This level of certainty is significant evidence of not meeting the standard for relative progress.
- **Well Below** is an indication that the relative progress measure is more than two standard errors below expected relative progress (0). This level of certainty is significant evidence of not meeting the standard for relative progress.

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

$$Relative\ Progress\ Index = \frac{Relative\ Progress\ Measure - Expected\ Relative\ Progress}{Standard\ Error\ of\ the\ Relative\ Progress\ Measure} \tag{21}$$

The relative progress 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 relative progress measure is decidedly above or below expected relative progress. The relative progress index is useful when comparing value-added measures from different assessments or in different units, such as NCEs or scale scores. The categories can be established as ranges based on the relative progress index, such as the following:

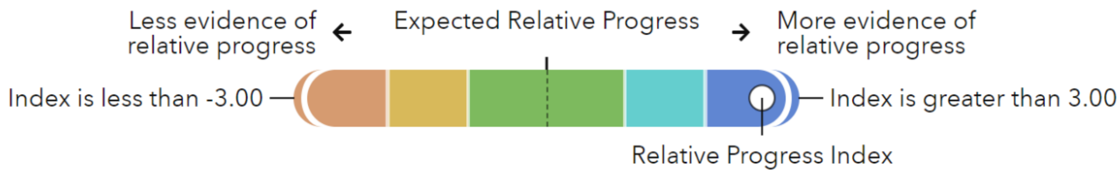
- **Exceeds Expected Relative Progress (Dark Blue)** indicates significant evidence that students made more relative progress than expected. The index is 2 or greater.

- **Meets Expected Relative Progress (Green)** indicates evidence that students made relative progress as expected. The index is between -2 and 2.

- **Does Not Meet Expected Relative Progress (Red)** indicates significant evidence that students made more relative progress than expected. The index is less than -2.

This is represented in the relative progress indicator bar in Figure 5, which is similar to what is provided in the reports in the EVAAS web application. The black dotted line represents expected relative progress. The color-coding within the bar indicates the range of values for the relative progress index within each category.

Figure 5: Sample Relative Progress Indicator Bar



4.5 Rounding and Truncating Rules

As described in the previous section, the effectiveness level is based on the value of the index. As additional clarification, the calculation of the index uses unrounded values for the value-added measures and standard errors. After the 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 the Idaho Relative Progress Model

5.1 Assessment Data Used in Idaho

For the analysis and reporting based on the 2021-22 school year, EVAAS receives the following assessments for use in the relative progress and/or projection models:

- IRI K-3 (Composite, Listening Comprehension, Letter Knowledge, Phonemic Awareness, Alphabetic Decoding, Reading Comprehension, Vocabulary, Spelling, and Text Fluency)
- ISAT English Language Arts and Mathematics in grades 3–8, 10
- ISAT Science in grades 5, 8, 11
- PSAT 8/9 and PSAT NMSQT Assessments in Evidence-Based Reading and Writing and Mathematics
- SAT Assessments in Evidence-Based Reading and Writing and Mathematics

Assessment files provide the following data for each student score:

- Scale score
- Performance level
- Test taken
- Tested grade
- Tested semester
- Administration window
- LEA number
- School number

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

5.2 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 progress. SAS receives this information in the form of various socioeconomic, demographic, and programmatic identifiers provided by OSBE. Currently, these categories are as follows:

- Gender (M, F)
- Students with Disabilities (Y, N)
- Students Who are Homeless (Y, N)
- Students from Migrant Families (Y, N)
- Students from Military Families (Y, N)
- Students in Foster Care (Y, N)
- Students Learning English (Y, N)
- Students Who are At Risk (Y, N)
- Students Who are Chronically Absent (Y, N)
- Students Who are Economically Disadvantaged (Y, N)
- Title I (School-level characteristic, Y, N)
- Students Who are Continuously Enrolled (School, District and State; District and State; State; Not Continuously Enrolled)

- Race
  - Asian
  - Black/African American
  - Hispanic or Latino
  - Multiracial
  - Native American or Alaskan Native
  - Native Hawaiian or Other Pacific Islander
  - White

6 Business Rules

6.1 Assessment Verification for Use in Relative Progress Models

To be used appropriately in any relative progress 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 relative progress model, and Idaho's current standardized assessments meet them. These criteria are explained in more detail below.

6.1.1 Stretch

Stretch indicates whether the scaling of the assessment permits relative progress to be measured for both very low- or very high-achieving students. A test "ceiling" or "floor" inhibits the ability to assess students' relative progress 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 relative progress at the very top of the scale when it is likely due to the artificial ceiling of the assessment. Percentages for all Idaho 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. The type of reliability is important for most any use of standardized assessments.

6.2 Pre-Analytic Processing

6.2.1 Missing Grade

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

6.2.2 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.

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

If a student has a score with a missing district or school for a particular subject and grade in a given testing period, then the duplicate score that has a district and/or school will be included over the score that has the missing data. If one record has more student demographic fields filled out, that record will be retained.

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

If a student has multiple scores in the same period for a particular subject and grade and the test scores are not the same, then those scores will be excluded from the analysis. If multiple SAT scores are present, only the closest test score to the average test date is kept. This is done by subject.

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

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

6.2.6 Students with Records That Have Unexpected Grade Level Changes

If a student skips more than one grade level (e.g., moves from sixth in 2018 to ninth in 2019) or is moved back by one grade or more (i.e. moves from fourth in 2018 to third in 2019) 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 kept in the analysis but only on cohort scores will be used.

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

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

6.2.8 Excluding High School ISAT Records Other Than Grade 10 from Analysis

Only high school ISAT records associated with grade 10 are included in the analysis. All other high school ISAT records associated with other grade levels are excluded.

6.2.9 Excluding SAT Records Other Than Grade 11 From Analysis

Only SAT records associated with grade 11 are included in the analysis. SAT records from all other grade levels are excluded.

6.2.10 PSAT and PSAT NMSQT Exclusion Rules

PSAT 8/9 and PSAT 10 records are not included in the analysis due to low counts. For PSAT NMSQT, only grade 10 records are included in the analysis.

6.2.11 Exclude Historical IRI Records

Test records for IRI from the 2017-18 school year and prior are excluded from analysis.

6.2.12 Normal Curve Equivalent (NCE) creation for IRI Composites

The NCE's created for IRI Composites used in the analysis are recentered so that they have mean of 50 and standard deviation of 21.06. This recentering only applies to IRI composites and not the component scores.

6.2.13 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 subjects are examined simultaneously, 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 assessments 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 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 relative progress 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 IRI and ISAT ELA and Math assessments or below -4.0 for ISAT Science, PSAT, and SAT assessments 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 or 4.0 standard deviations below the comparison score depending on the assessment.
- 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 IRI and ISAT ELA and Math assessments or above 5.0 for ISAT Science, PSAT, and SAT assessments 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 or 5.0 standard deviations above the comparison score depending on the assessment.
- 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.

6.2.14 Linking Records over Time

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

6.3 Relative Progress Models

6.3.1 Students Included in the Analysis

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

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

For the predictive and projection models, a student must have at least three valid predictor scores that can be used in the analysis, all of which cannot be deemed outliers. (See Section 6.2.11 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 relative progress. The required three predictor scores are needed to sufficiently dampen the error of measurement in the tests to provide a reliable measure. If a student does not meet the three-score minimum, then that student is excluded from the analyses. It is important to note that not all students have to have the same three prior test scores; they only have to have some subset of three that were used in the analysis. Unlike the gain model, students with non-traditional grade patterns are included in the predictive model as one student. Since the predictive model does not determine relative progress based on consecutive grade movement on tests, students do not need to stay in one cohort from one year to the next. That said, if a student is retained and retakes the same test, then that prior score on the same test will not be used as a predictor for the same test as a response in the predictive model. This is mainly due to the fact that very few students used in the models have a prior score on the same test that could be used as a predictor. In fact, in the predictive model, it is typically the case that a prior test is only considered a possible predictor when at least 50% of the students used in that model have those prior test scores. Students are excluded from the predictive model if the student is flagged as a First Year EL student or if the student does not meet partial enrollment membership for IRI and ISAT assessments. There are no membership rules used to include or exclude students in the SAT or PSAT analyses.

6.3.2 Minimum Number of Students to Receive a Report

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

6.3.2.1 District and School Model

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

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

subject. However, due to the rule above, it is typically the case that at least five students have a "simple gain." In some cases where students only have a Math or Reading score in the current year or previous year, this value dips below five.

- For any district or school relative progress measures based on specific student groups, the same requirements described above apply for the students in that specific student group.

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

For any district or school relative progress measures based on specific student groups, the same requirements described above apply for the students in that specific student group.