EDUCATIONAL DATA SYSTEMS

# EdScale

## How to Measure Growth using Formative Exams Without Common Persons or Items

**Mark H. Moulton, Ph.D.**

**March 11, 2010**

A new method is proposed for equating multidimensional formative exams without common persons or items. A multidimensional IRT (MIRT) model called NOUS is used to link students who take different formative exams by exploiting scores received on a common test administered at some point in the recent past, such as the California Standards Test. The “past test” vector is projected into the multidimensional subspace of the two formative exams, and students located in the same subspace are projected onto the common “past test” vector, allowing apples-to-apples comparisons between students on a common, well-understood metric. Growth measurement is handled by applying a time-series function to expected growth rates computed from previous years. The methodology is presented in connection with a scaling product developed in California, called EdScale, which is used to measure student growth on benchmark exams developed or purchased independently by school districts.

# EdScale and Formative Assessment

## Purpose

Evidence is accumulating that formative assessment, properly used, is a cost-effective way to increase student achievement (Wiliam & Leahy, 2009). Formative assessment is defined as the extent that evidence about student achievement is elicited, interpreted, and used by teachers to make decisions about the next steps in instruction. However, formative assessments present serious psychometric challenges. As their purpose is diagnostic rather than summative, formative assessments are often built, quite properly, to be multidimensional; they contain multiple content areas to achieve a more holistic view of student abilities. However, without a single well-designed construct, test equating appears to become untenable. Students who take one formative assessment cannot be compared to students who take another, and they cannot be compared to themselves over time to measure growth.

We thus see a psychometric tension between formative and summative assessment. A properly equated summative assessment permits comparison of students across tests and over time, but at the expense of useful diagnostic information at a point in time. On the other hand, a diagnostically useful formative assessment lacks the properties needed to provide valid student comparisons across tests. We can construct a good formative test or a good summative test, but apparently not both at the same time. That is the dilemma.

There have been a number of serious efforts to negotiate the formative/summative dilemma. The simplest and most widely used is to do both kinds of testing – formative through the school-year, one summative exam at the end of the year. The difficulty is that one summative end of year test does not permit measurement of within-grade student-level growth and most formative or benchmark exams are not able to fill the void. Meanwhile, many formative exams are not even particularly useful for their ostensible purpose of providing actionable student diagnostics and guiding instructional decisions.

At the other extreme, companies like CTB/McGraw-Hill (Lewis, 2009) and NWEA (Kingsbury, 2005) offer pre-equated benchmark testing systems which, though perhaps not ideal as formative assessments, are nonetheless effective and allow teachers to track progress of their students through the school year while capturing useful diagnostic information. However, such systems often require districts to purchase a pre-equated item bank, or spend time creating one, and there is a cost in flexibility. Also, they may not be well aligned to individual district pacing schedules.

A third approach to negotiating the formative/summative dilemma is described in this paper. It entails a new method of test equating called "common history equating". This approach draws on Multidimensional Item Response Theory (MIRT), in particular a model implemented using a Rasch-like matrix decomposition methodology called NOUS. NOUS, in turn, is part of a larger scaling package known commercially as EdScale, offered by Educational Data Systems, Inc., in Morgan Hill, California. As of the date of this paper, EdScale is implemented as a feature of the STARS reporting platform hosted by School City in Mountain View, California. Districts load their benchmark test data onto STARS™ and can, within a day, view their test scores on a CST-equivalent scale, with state-aligned performance levels. These scores answer the question, "What would each student have received on the CST (Math or Language) if they had taken the CST in place of the benchmark test?"

Thus, EdScale makes it possible to do the following:

- Compute actionable and reliable student diagnostic statistics
- Place students who take different formative assessments on a common Math or Language CST scale, without common persons or items
- Measure student growth
- Forecast end-of-year CST scores
- Automatically assign valid state-aligned performance levels
- Improve the quality and reliability of their tests

It does these things without requiring districts to purchase tests from any particular vendor and without a lengthy preparation period. Tests can be changed or replaced from year to year without compromising the scaling system. No standard setting meetings are necessary to set cut-scores, as they are automatically equated to the state's performance levels. "Proficient" on the benchmark test means the same thing as "Proficient" on the CST for each grade.

The method of computing student diagnostic statistics is not discussed here. What is discussed is the method of placing students on a common scale – common history equating – and how this is used in connection with other information to measure student growth through the school year and across grades.

## The Underlying Methodology: NOUS

EdScale rests on NOUS, a flavor of Multidimensional Item Response Theory (MIRT) (Moulton, 2010).[1] NOUS employs an algorithm, called Silsdorf's decomposition after its inventor that decomposes a multidimensional data matrix **X**, which can contain missing data, into a row coordinates matrix **R** and a column coordinates matrix **C** with a specified number of dimensions *D*. Each datum is specified to be the product of a row entity interacting with a column entity. The Euclidean inner product of **R** and **C** produces a matrix of estimates **E**:

$$\mathbf{E} = \mathbf{RC} \qquad \text{Eq. 1}$$

[1] See Reckase, 2009, for a history and review of Multidimensional Item Response Theory.

When **E** is obtained through an iterative process of alternating least squares solutions, it can be shown that **E** is the orthogonal projection of **X** into a subspace of *D* dimensions (Moulton, 2010). Silsdorf's decomposition yields estimates that approximately equal those obtained by other matrix decomposition techniques, such as Singular Value Decomposition (SVD). However, the NOUS decomposition has several useful properties not shared by SVD:

1. Robust to Missing Data. Like the Rasch Model, the NOUS decomposition algorithm computes coordinates for each row and column one entity at a time, with whatever data is available for that entity (Wright, 1982). This makes it straightforward to analyze extremely sparse data matrices and predict missing cells as necessary. No assumptions need be made about the nature of the missing data.
2. Objective Dimensionality Criterion. Because the model is robust to missing data, the NOUS algorithm calculates the optimal dimensionality for the data matrix by making cell values randomly missing and assessing the model's prediction error for each of a set of possible dimensionalities. In a dataset that meets the specification of the model, the "true" or "objective" dimensionality is the one that yields cell predictions that most closely match the observed cell values. This will be the minima of a U-shaped prediction error curve, where the y-axis is the error and the x-axis is the number of dimensions used to compute predictions.
3. Anchoring. The coordinates calculated from one data set can be used to anchor the rows or columns of another data set, forcing both data sets to share the same coordinate space. This is exactly analogous to Rasch anchoring.

Once an optimal dimensionality has been determined, coordinates are calculated for each row and column entity at that number of dimensions. In educational psychometrics, the row entities are generally "persons" or "students" and the column entities are "items" or "questions". Estimates and observed values are compared through analysis of fit and entities deleted where necessary. NOUS analysis of fit is in principle no different from Rasch analysis of fit. The result is a set of matrix estimates that meet Rasch criteria for reproducibility across data sets. When the data fit the model, each estimate of person performance on an item will in principle be the same regardless of the sample of items and persons used to calculate it, so long as they reside in the same space. Note, however, that unlike Rasch ability and difficulty parameters, NOUS coordinates are largely arbitrary and uninterpretable; test measures and predictions are drawn from the estimates matrix **E**.

In the special case where $D = 1$, Equation 1 is similar to a non-probabilistic form of Rasch model. Performance on an item is the product (or log sum) of a person "ability" parameter and an item "easiness" parameter (the inverse of "difficulty"). NOUS models exist for dichotomous, polytomous, interval, and ratio data.

## The Common History Equating Problem

The problem that common history equating tries to solve is this:  how do we compare students who take different formative tests that may have no common items, by using their scores on a third test administered in the recent past?

Let there be two formative tests, Test 1 and Test 2, with no items in common.  Let us assume they have been administered on the same date, in October, to two different samples of students.  Assume that the test items are individually or collectively sensitive to more than one content or dimension, but that the two tests do not have the same mix of contents.  For example, let us imagine that Test 1 is weighted toward Grammar 75% and Comprehension 25%, while Test 2 is weighted 25% Grammar and 75% Comprehension.  How do we compare the students who took Test 1 to the students who took Test 2?

Traditionally, there are two equating strategies:  link the tests using common items or common persons.  In this typical, real-world situation, neither strategy is available.  The students only take one test, and there are no linking items between the two tests.  Even with linking items, the lack of a single content dimension would scuttle efforts to apply a standard Rasch equating model.

However, let us assume that all students took a "state exam" the previous May.  Is there a way to use this information somehow to link and compare the two sets of students?  Yes, and this is the method discussed in this paper under the name "common history equating".

## A Simulated Data Set

To explore this question, we first need to construct a data set that embodies the common history problem and that we can use to test the accuracy of any methodologies we come up with.  Figures 1 and 2 offer such a data set.  They are artificially generated data matrices for two hypothetical tests, Test 1 and Test 2.  We imagine in this case that the same classroom of students has taken both tests at some date in October and that Test 1 is weighted in favor of Content X (such as Grammar) and Test 2 is weighted in favor of Content Y (such as Comprehension).

We suppose, further, that the students took a third test, a state test like the California Standards Test, in May of the previous school year.  Figures 1 and 2 show what the resulting data matrices might look like.

Figure 1: Test 2 Simulation, October item scores with scores on "state test" administered last May

| | Test 1: Weighted toward Content X, | | | | | | | | | State | Common |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | administered in October | | | | | | | | | Test | Scale |
| Pers/It | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | May | Oct |
| Abe | 7 | 13 | 5 | 11 | 2 | 9 | 12 | 2 | 4 | 1 | 66 | ? |
| Betty | 2 | 4 | 2 | 3 | 1 | 2 | 3 | 1 | 1 | 1 | 11 | ? |
| Chia | 2 | 4 | 2 | 4 | 1 | 2 | 3 | 2 | 1 | 1 | 54 | ? |
| Dale | 5 | 9 | 4 | 7 | 1 | 7 | 9 | 1 | 3 | 1 | 71 | ? |
| Eric | 9 | 17 | 7 | 14 | 2 | 12 | 15 | 3 | 5 | 2 | 159 | ? |
| Flo | 6 | 11 | 5 | 10 | 2 | 7 | 10 | 2 | 4 | 1 | 116 | ? |
| Gloria | 2 | 5 | 3 | 4 | 1 | 1 | 2 | 2 | 2 | 1 | 89 | ? |
| Han | 1 | 2 | 2 | 2 | 0 | 1 | 1 | 1 | 1 | 1 | 92 | ? |
| Ian | 5 | 10 | 5 | 9 | 2 | 6 | 9 | 2 | 3 | 1 | 102 | ? |
| Juju | 8 | 14 | 6 | 12 | 2 | 10 | 13 | 2 | 4 | 1 | 80 | ? |
| Kara | 8 | 13 | 5 | 11 | 2 | 10 | 13 | 2 | 4 | 1 | 101 | ? |
| Li | 6 | 10 | 5 | 9 | 2 | 7 | 9 | 2 | 3 | 1 | 80 | ? |
| Maria | 1 | 3 | 2 | 2 | 0 | 1 | 2 | 1 | 1 | 0 | 18 | ? |
| Ned | 10 | 18 | 8 | 16 | 3 | 12 | 16 | 4 | 6 | 2 | 198 | ? |
| Oz | 5 | 11 | 6 | 10 | 2 | 6 | 9 | 3 | 4 | 1 | 134 | ? |
| Pedro | 8 | 14 | 6 | 12 | 2 | 10 | 13 | 2 | 4 | 1 | 168 | ? |
| Qing | 6 | 11 | 5 | 9 | 2 | 7 | 10 | 2 | 3 | 1 | 67 | ? |
| Rolf | 5 | 9 | 4 | 8 | 1 | 6 | 8 | 2 | 3 | 1 | 107 | ? |
| Satya | 5 | 9 | 4 | 8 | 1 | 6 | 8 | 2 | 3 | 1 | 105 | ? |
| Truth | 6 | 10 | 5 | 9 | 2 | 7 | 9 | 2 | 3 | 1 | 98 | ? |

## Notes

- Each score for items 1 – 10 is the product of a person vector and an item vector. Each vector is 2-dimensional, i.e., has an X component and a Y component in an X, Y coordinate system. The person vectors were randomly generated. The item vectors were also randomly generated, except that the first dimension X was adjusted to have larger values than the Y dimension for all items.

- The "State Test" vector represents scores on a test administered in May of the previous school year. It was simulated by adding the item scores across the row for both Tests 1 and 2, then subtracting a random number between 0 and 60. This represents the idea that the State Test has its own mix of X and Y contents which are a compromise between the content weights assigned to Test 1 and Test 2. The random number represents the random variation caused by students having different growth rates from May to October, as well as random variation caused by the inclusion of additional content areas on the State Test that are not included on either Test 1 or Test 2.

Figure 2: Test 2 Simulation, October item scores with scores on “state test” administered last May

| | Test 2: Weighted toward Content Y, | | | | | | | | | | State | Common |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | administered in October | | | | | | | | | | Test | Scale |
| **Pers/It** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **May** | **Oct** |
| **Abe** | 6 | 3 | 5 | 2 | 5 | 6 | 5 | 4 | 2 | 4 | 66 | ? |
| **Betty** | 5 | 4 | 5 | 2 | 6 | 7 | 3 | 3 | 2 | 2 | 11 | ? |
| **Chia** | 6 | 6 | 7 | 3 | 8 | 10 | 5 | 5 | 2 | 2 | 54 | ? |
| **Dale** | 2 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 71 | ? |
| **Eric** | 9 | 6 | 8 | 4 | 8 | 11 | 7 | 6 | 3 | 5 | 159 | ? |
| **Flo** | 9 | 7 | 9 | 4 | 10 | 12 | 7 | 6 | 3 | 4 | 116 | ? |
| **Gloria** | 9 | 9 | 9 | 4 | 12 | 14 | 7 | 7 | 3 | 3 | 89 | ? |
| **Han** | 5 | 5 | 5 | 3 | 7 | 8 | 4 | 4 | 2 | 1 | 92 | ? |
| **Ian** | 8 | 6 | 8 | 4 | 9 | 11 | 6 | 6 | 3 | 3 | 102 | ? |
| **Juju** | 7 | 4 | 6 | 3 | 6 | 8 | 5 | 4 | 2 | 4 | 80 | ? |
| **Kara** | 4 | 2 | 4 | 2 | 2 | 4 | 4 | 3 | 1 | 3 | 101 | ? |
| **Li** | 6 | 4 | 5 | 2 | 5 | 7 | 5 | 4 | 2 | 3 | 80 | ? |
| **Maria** | 4 | 4 | 4 | 2 | 5 | 6 | 3 | 3 | 1 | 1 | 18 | ? |
| **Ned** | 12 | 9 | 12 | 5 | 13 | 16 | 10 | 8 | 4 | 6 | 198 | ? |
| **Oz** | 10 | 9 | 10 | 5 | 12 | 15 | 8 | 7 | 4 | 4 | 134 | ? |
| **Pedro** | 8 | 5 | 7 | 3 | 7 | 9 | 6 | 5 | 3 | 4 | 168 | ? |
| **Qing** | 6 | 4 | 5 | 2 | 5 | 7 | 5 | 4 | 2 | 3 | 67 | ? |
| **Rolf** | 5 | 3 | 5 | 2 | 5 | 6 | 4 | 3 | 2 | 3 | 107 | ? |
| **Satya** | 6 | 4 | 5 | 3 | 6 | 7 | 4 | 4 | 2 | 3 | 105 | ? |
| **Truth** | 5 | 4 | 5 | 2 | 5 | 7 | 4 | 4 | 2 | 3 | 98 | ? |

## Notes

- Each score for items 11 – 20 is the product of a person vector and an item vector. Each vector is 2-dimensional, i.e., has an X component and a Y component in an X, Y coordinate system. The person vectors were randomly generated. The item vectors were also randomly generated, except that the second dimension Y was adjusted to have larger values than the X dimension for all items.

- The “State Test” vector represents scores on a test administered in May of the previous school year. It was simulated by adding the item scores across the row for both Tests 1 and 2, then subtracting a random number between 0 and 60. This represents the idea that the State Test has its own mix of X and Y contents which are a compromise between the content weights assigned to Test 1 and Test 2. The random number represents the random variation caused by students having different growth rates from May to October, as well as random variation caused by the inclusion of additional content areas on the State Test that are not included on either Test 1 or Test 2.

The Notes explain how each item score was generated. Person vectors and item vectors were randomly generated for a 2-dimensional coordinate system, where the first dimension is called X and the second dimension is called Y. In test 1, the X dimension of the item vectors is adjusted to be larger than the Y dimension, and vice versa for Test 2. This simulates the idea that Test 1 is weighted in favor of the X dimension (e.g., Grammar) and Test 2 in favor of the Y dimension (e.g., Comprehension). Each score is the dot product of the person vector for that row and the item vector for that column.

Each score differs from what one would expect in the real world in two ways. First, test data is usually dichotomous (0, 1) instead of interval. Second, real-world data is subject to large amounts of measurement error, random variation in the observed data that has nothing to do with a student's actual abilities. For purposes of this simulation, we need to be able to assume that the test data for Test 1 and Test 2 have no measurement error so that we can focus on the effects of error in the "State test" variable. That way, when we measure the efficacy of the proposed equating methodology described below, we can do so without having to worry about the confounding effects of random measurement error in the tests. If our proposed equating methodology works, and there is no measurement error in Tests 1 and 2, then students should receive exactly the same equated common scale scores on the two tests.

To test the equating methodology proposed below, we have made the vector coordinates for each person the same for Tests 1 and 2, because the persons are the same across the two tests. This simulates the idea that students have the same underlying abilities regardless of which test they take and provides a method for testing the efficacy of the proposed equating methodology.

As described in the Notes, the "State Test" variable is simulated by summing the scores across the rows for the two tests, then subtracting a random number between 0 and 60 for each row. By summing the scores across the two tests, we simulate the idea that the State Test contains at least the two content areas as Tests 1 and 2 (Grammar and Comprehension), but weighted differently and more evenly. It doesn't matter how we combine the two tests to simulate the state test, just that the state test contains some information regarding the Test 1 and Test 2 dimensions.

Because the state test is assumed to have been administered 6 months before Tests 1 and 2, and because we can assume that it contains additional content areas besides X and Y (Grammar and Comprehension), we introduce a large random element. This is done by subtracting a random number between 0 and 60 from the sum of row scores for each person, quite a substantial perturbation given that the range of simulated test scores without the perturbation is 142 points. This simulates a combination of "trend-variation" (different growth rates for each student between May and October) and "content-variation" (the effect of extra contents that are not included in Tests 1 and 2). The task of the common history equating methodology described below, therefore, is somehow to strip away the effects of trend-variation and content variation, so that students are compared only in terms of their performance on the formative tests.

The final column in Figures 1 and 2 represents the common scale score that we want for Tests 1 and 2 but do not have, a scale score that allows us to compare students who take Test 1 with students who take Test 2. Because, for purposes of this simulation, we have hypothesized that the same students take both tests, we ask the following question:

*Given the common history equating methodology proposed below, is it possible to compute common scale scores for a given student on Test 1 that match the common scale scores computed for the same student based on data in Test 2?*

An affirmative answer to this question would establish the theoretical basis of the equating methodology proposed below.

## Common History Equating Methodology

The procedure proposed for equating two tests in terms of common "past test" scores is as follows:

1. Find $\mathbf{R}_1$. Run Test 1 data through Silsdorf's decomposition to obtain $N$ Test 1 person coordinates $\mathbf{R}_1$ in a 2-dimensional space ($\mathbf{R}_1$ is an $N$ x $D$ array, where $D = 2$), spanning the two content dimensions X and Y. (The number of dimensions is found empirically by successively decomposing the matrix at different dimensionalities and observing which number of dimensions yields estimates that best predict pseudo-missing cells.)

2. Find $\mathbf{R}_2$. Likewise, decompose Test 2 to obtain an $N$ x $D$ array of Test 2 person coordinates $\mathbf{R}_2$. Tests 1 and 2 are assigned the same number of dimensions. Note that the Test 1 and Test 2 data used to calculate **R** do not include the "state exam" scores.

3. Find $\mathbf{C}_1$ and $\mathbf{C}_2$ for the "state exam". Use the columns in coordinates array $\mathbf{R}_1$ as independent variables to predict the state exam scores administered last May. The Gaussian least squares solution for $\mathbf{C}_1$ is:

$$\mathbf{C}_1[\textit{solution}] = (\mathbf{R}_1^{\mathbf{T}}\mathbf{R}_1)^{-1}\,\mathbf{R}_1^{\mathbf{T}}\,\mathbf{X} \qquad \text{Eq. 2}$$

   where **X** is the column of state exam scores. $\mathbf{C}_2[\textit{solution}]$ is calculated the same way.

4. Compute estimates $\mathbf{E}_1$ of how each person would perform on the "state exam" based only on their performance on Test 1 (or Test 2) as encapsulated in their **R** person coordinates.

$$\mathbf{E}_1 = \mathbf{R}_1\mathbf{C}_1 \qquad \text{Eq. 3a}$$
$$\mathbf{E}_2 = \mathbf{R}_2\mathbf{C}_2 \qquad \text{Eq. 3b}$$

$\mathbf{E}_1$ and $\mathbf{E}_2$ are equated person measures that place the two sets of persons on a common scale defined by the metric of the state exam after filtering out all dimensions extraneous to Test 1 and Test 2. This "filtering out" property follows from a very important theorem of linear least squares discovered by Gauss in the 1790's according to which $\mathbf{E}$ can be interpreted as the orthogonal projection of state exam scores $\mathbf{X}$ into the $D$-space erected independently by Tests 1 and 2, embodied as the $\mathbf{R}_1$ and $\mathbf{R}_2$ coordinate systems (Wikipedia). Stated plainly, we can equate two tests to the degree we can use their data to predict the same variable.

Between the spring state exam and the fall benchmark exam, we expect that students will have grown at different rates. These differential growth rates, as well as other sources of variation relating to the state exam, are automatically filtered out by projecting the original state exam scores into the sub-space defined by Tests 1 and 2. The resulting projections $\mathbf{E}_1$ and $\mathbf{E}_2$ answer the question: How would each student have performed *relative to the others* on the state exam if the state exam had been administered in October rather than May.

I emphasize "relative to the others" because the common history procedure does not account for the overall increase in mean ability from May to October, the least squares estimates being constrained to have the same mean as the original observations. Therefore, this procedure cannot on its own be used to measure growth and for purposes of the simulation we must assume that Tests 1 and 2 were administered at the same time. To measure growth, we need additional steps discussed in a later section.

## Simulation Results

We now apply the common history equating procedure described above to the simulated data sets in Figures 1 and 2.

Figure 3 compares the person raw scores obtained on the two tests. It confirms that even though both tests have the same X and Y content, the different weightings make their raw scores impossible to compare. The total raw scores on Test 1 bear little relation to those on Test 2, even though the underlying contents are the same and even though the same persons are taking both tests. This shows that unidimensional IRT models such as the Rasch Model would not be effective here.

Figure 4 shows what we get after running Test 1 and Test 2 through the common history procedure above. The two columns contain scale score measures separately generated from the data in Test 1 and Test 2. We see that they are virtually identical and that the perturbations caused by trend-variation and content-variation have been filtered out. The two tests have been successfully equated.

Figure 3: Test 1 Raw Scores vs. Test 2 Raw Scores, r = 0.36

| | Raw Scores | |
|---|---|---|
| | Test 1 | Test 2 |
| Abe | 66 | 41 |
| Betty | 20 | 38 |
| Chia | 21 | 54 |
| Dale | 46 | 16 |
| Eric | 87 | 67 |
| Flo | 59 | 70 |
| Gloria | 22 | 76 |
| Han | 11 | 43 |
| Ian | 51 | 63 |
| Juju | 73 | 49 |
| Kara | 69 | 29 |
| Li | 52 | 43 |
| Maria | 13 | 34 |
| Ned | 94 | 94 |
| Oz | 57 | 84 |
| Pedro | 74 | 58 |
| Qing | 55 | 43 |
| Rolf | 47 | 38 |
| Satya | 47 | 44 |
| Truth | 53 | 41 |

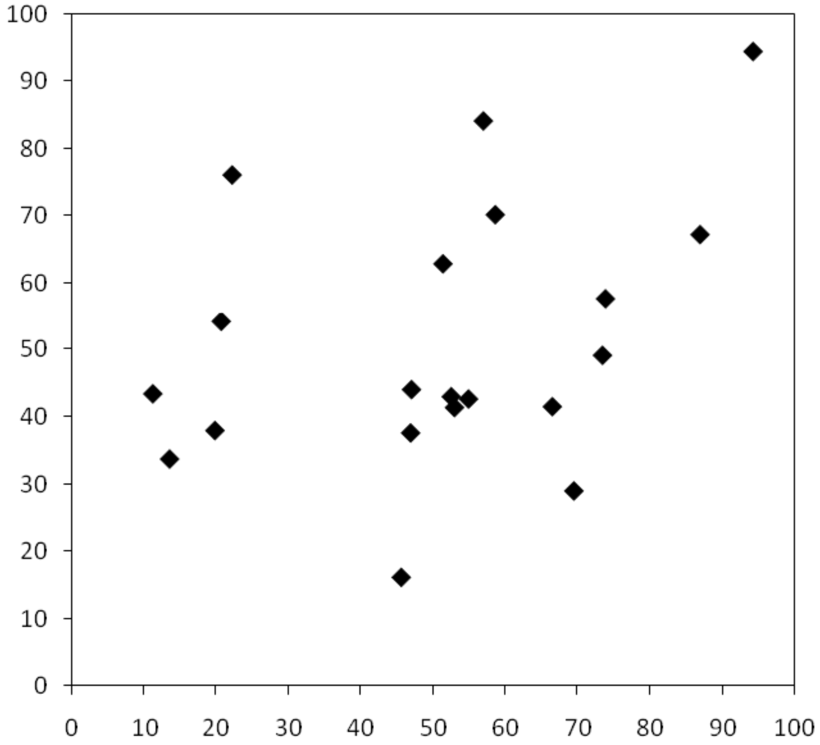
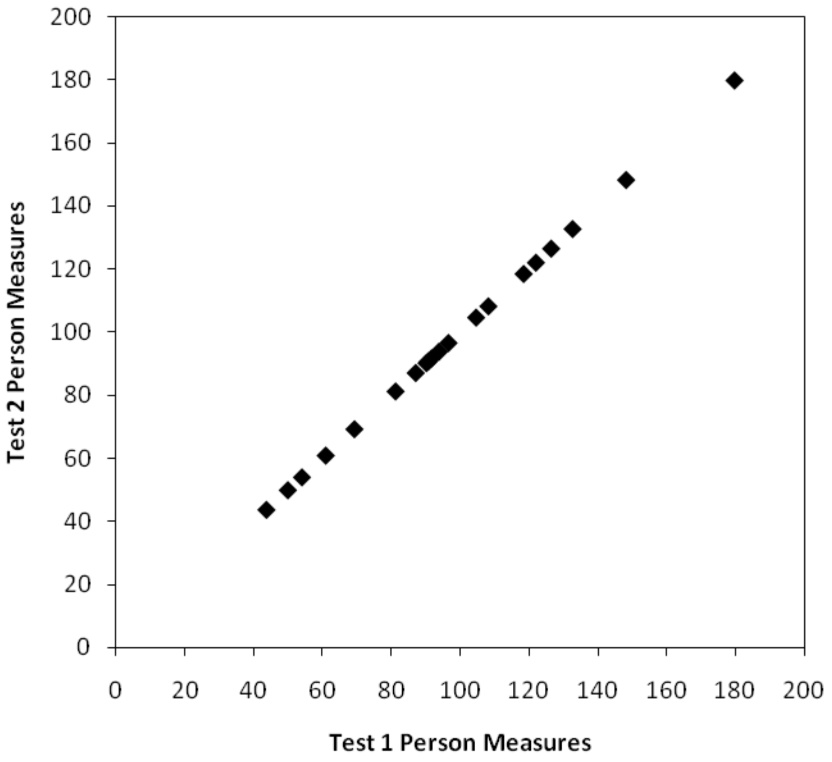


Figure 4: Test 1 vs. Test 2 Common Scale Equated Measures, October, r = 1.00

| | Common Scale Equated Measures | |
|---|---|---|
| | Test 1 | Test 2 |
| Abe | 104 | 104 |
| Betty | 54 | 54 |
| Chia | 69 | 69 |
| Dale | 61 | 61 |
| Eric | 148 | 148 |
| Flo | 122 | 122 |
| Gloria | 90 | 90 |
| Han | 50 | 50 |
| Ian | 108 | 108 |
| Juju | 118 | 118 |
| Kara | 96 | 96 |
| Li | 92 | 92 |
| Maria | 43 | 43 |
| Ned | 180 | 180 |
| Oz | 133 | 133 |
| Pedro | 126 | 126 |
| Qing | 94 | 94 |
| Rolf | 81 | 81 |
| Satya | 87 | 87 |
| Truth | 91 | 91 |



Figures 3 and 4 show that by exploiting the principle of orthogonal projection implicit in Gaussian least squares, we can compare students who take different formative or benchmark assessments in terms of their performance on a third test, such as a state exam, administered at an earlier date. The projection process filters out temporal and content differentiation effects.

The straight line relationship in Figure 4 follows from: a) the principle that a given vector (calculated from the state exam scores) projected onto two subspaces will yield identical estimates so long as the two subspaces span the same content dimensions; and b) the fact that $\mathbf{R}_1$ and $\mathbf{R}_2$ are error-free due to the test response data $\mathbf{X}_1$ and $\mathbf{X}_2$ being error-free. Introducing measurement error to $\mathbf{X}_1$ and $\mathbf{X}_2$ (adding random numbers to the score data) would cause the straight line relationship in Figure 5 to become less precise and devolve into a cigar shape. The relationship between the addition of measurement error and the resulting imprecision in $\mathbf{E}$ is governed by the Gauss-Markov theorem, which states that the validity of a linear least squares solution is a function of the following factors: a) the extent of the error added; b) the number of test items; c) whether the test measurement errors are homoscedastic (the same on average across the data); d) whether they sum to zero; and e) whether they are uncorrelated. A full empirical study of the degree to which real-world educational test data fail to meet these requirements, thus weakening the comparability of different tests, falls outside the scope of this paper, though an empirical case study is provided in this paper.

What is important is the following remarkable property:

*The ability to equate two tests in terms of a third test administered in the past is determined by factors largely within the control of the developers of the two tests to be equated.*

These factors are: a) minimizing test measurement error (e.g., maximize the number of items); b) ensuring content homogeneity (write items to get at the same underlying content dimensions for both tests, albeit with different weights); c) ensuring that both tests contain the same underlying content dimensions as those included on the third test administered in the past (though additional content dimensions are permitted since these will be filtered out along with the trend-variation).

Given these constraints, it is possible to describe the circumstances under which common history equating is likely to be successful. For instance, we can expect that the precision of the equating procedure will decline the longer the time elapsed since the common test was administered, and that the rate of decline will be a function of the measurement error of the test and to a lesser extent the trend-noise that has occurred since the test was administered. A theoretically complete function relating equating error to time elapsed remains to be specified, however.

## Measuring Growth

So far, common history equating has been described as a way to compare students who take two different tests *at the same time*. What if the tests are administered at *different* times, for instance in

October and December? A more elaborate procedure must be used. In the EdScale algorithm, the procedure is as follows:

1. Equate the State Tests. California administers the California Standards Test (CST) for Grades 2 – 11. While the tests are equated across time, so that grade 5 students in 2009 can be compared to grade 5 students in 2010, they are not equated across grades. Grade 5 students can't be compared to grade 4 students. The EdScale strategy for measuring growth requires that student scores be comparable across grades. Therefore, Educational Data Systems (the developers of EdScale) conducted an equating study to place the Math and Language CSTs on common cross-grade scales. For other states, EdScale relies on other equating studies such as those published by NWEA (see NWEA reference). Note that this has only been done for Math and Language. These are the only two academic content areas with sufficiently stable cross-grade constructs to allow for the measurement of cross-grade growth.

2. Convert Previous-Year CSTs to Common Scale. Let's say it is October and we want to scale an October benchmark exam. CST scores for Math and Language are collected for each student for the CST that was administered in the previous spring. Using the results of the cross-grade CST equating study, we convert the CST scores into scores on the common scales for Math and Language. Now all the students across grades are on the same scale.

3. Compute Expected Growth Rates. Using the district's CST file for all grades, we calculate the mean and standard deviation of the students in each grade on the common scale metric. That makes it possible to calculate the average cross-grade growth for each pair of adjacent grades as it occurred last year. Statistical interpolation is used to estimate the grade 1 to grade 2 growth rate.

4. Scale the Benchmark Exam. We apply the common history equating technique described above to project student responses on the October benchmark exam onto the CST metric defined by the CST given in the previous spring. All differential trend effects caused by different student growth rates over the intervening months are automatically filtered out. The scores are standardized, with mean of zero and standard deviation of one.

5. Forecast End-of-Year Mean and Standard Deviation. We know the mean and standard deviation of the CST scores students in the current grade received last year, and have converted them to the common scale. We have their expected growth rate for this grade, from Step 3. From these two pieces of information, we forecast the mean and standard deviation of the common scale scores we expect when students take the CST at the end of the current school year.

6. Estimate Current Mean and Standard Deviation. From Step 2 we have the mean and standard deviation of the students from the previous school year. From Step 5 we have a forecast of the mean and standard deviation that we expect from these students at the end of

this year. To this we add our knowledge of the administration date of the current benchmark exam. We calculate the percentage of the school year that has elapsed since the CST was given in the previous year, and apply this percentage to the expected yearly growth rate to estimate the mean and standard deviation we expect now, in October.

7. Convert to Current Grade CST Metric. Steps 2, 3, 5 and 6 are all conducted in the common scale metric. This now needs to be converted back into the CST metric of the *current* grade (not the previous grade).

8. Estimate Current CST Scores. Convert the student standardized benchmark scale scores from Step 4 into CST scale scores by multiplying each student's score by the estimated current CST standard deviation and adding the product to the estimated current CST mean. Now the benchmark scores are in the current grade CST metric, and we have an answer to the question, "What would each student have obtained on the CST if the CST had been administered in place of the benchmark exam?"

9. Repeat for Next Benchmark Exam. The same procedure is followed when a benchmark exam is administered in the following December. The new administration date allows us to update the percentage of expected yearly growth that has transpired since October, allowing us to estimate a new mean and standard deviation, which is then applied to the individual standardized benchmark scores as projected onto the CST metric. The December scale scores are now comparable to the October scale scores, and show each student's growth over the intervening months. Growth rates will differ across students depending on how they shifted their relative positions in the overall student ability distribution.

The procedure relies on several assumptions. It is assumed that the average growth rate of the current cohort of students for a given grade will be similar to that of the previous cohort of students. In general the assumption holds up well, but there are exceptions. Districts may undergo rapid change across successive years, affecting the relative growth rates of different student cohorts. There may be demographic shifts, or district reorganizations, or loss of funding. Another possibility is that a district may administer the test to a subset of schools that has a different growth rate than the district average. Such factors will cause the end-of-year forecast to be off, as well as the steepness of the predicted growth rate. What the assumption buys is the ability to scale tests right out of the box, without a previous year's worth of testing data for that test. If it turns out the end-of-year forecast was wrong, it is not difficult to rescale the previous year's tests to maximize accuracy.

Another assumption is that academic growth within the school year is linear. While the assumption is probably not strictly valid, especially in the lower grades, from a reporting perspective it makes little difference whether the growth trends are straight or curved.

In higher grade Mathematics – Algebra I, Geometry, Algebra II – additional assumptions and procedures are needed which fall outside the scope of this paper.

So far, we have only discussed within-grade growth. Because we are working on a cross-grade scale, it is trivial to report cross-grade growth as well. As yet, that feature is not implemented in STARS in order to reduce complexity.

## A Case Study

EdScale has been available on the STAR platform for two years and has serviced four California districts, plus a few others. One of those districts, District A, requested scaling for three benchmark exams per year in English Language Arts for grades 2 – 6. The benchmarks were administered in October, February, and May of the 2008-09 school year to approximately 5,400 students per exam. The EdScale results of the October 2008 exam are presented here and compared with the results of the ELA CST administered in May 2009, some 7 months after the October benchmark was scaled. While no individual case study is sufficient establish the validity of a methodology, District A's October benchmark results are fairly typical across grades, contents, and other districts and provide a good sense of how the methodology works.

Language tests were administered to 5,452 students in grades 2 – 6 on October 31, 2008, though scaled in December. The tests consisted of 55 reading and writing items, of which approximately 15 per test were deemed of poor quality on psychometric grounds. Alignment to California standards is unknown. Overall test reliability exceeded 90% in all cases.

For each student, three CST-equivalent scale scores were computed and performance levels assigned according to California Department of Education published cut-points. The three sets of scores were:

1. Previous Year CST-Equivalent Scale Score. This is the scale score that a student received in spring of the student's previous grade, converted to the metric of the current grade. So, if the student is currently in grade 5, this is her grade 4 CST score converted to the grade 5 CST metric. This conversion makes it possible to measure growth starting from the previous grade.

2. Current CST-Equivalent Scale Score. This is the scale score that a student is expected to receive if taking the CST in place of the benchmark exam.

3. Forecasted End-of-Year Scale Score. This is the scale score that a student is forecasted to receive at the end of the year when she takes the CST corresponding to the current grade.

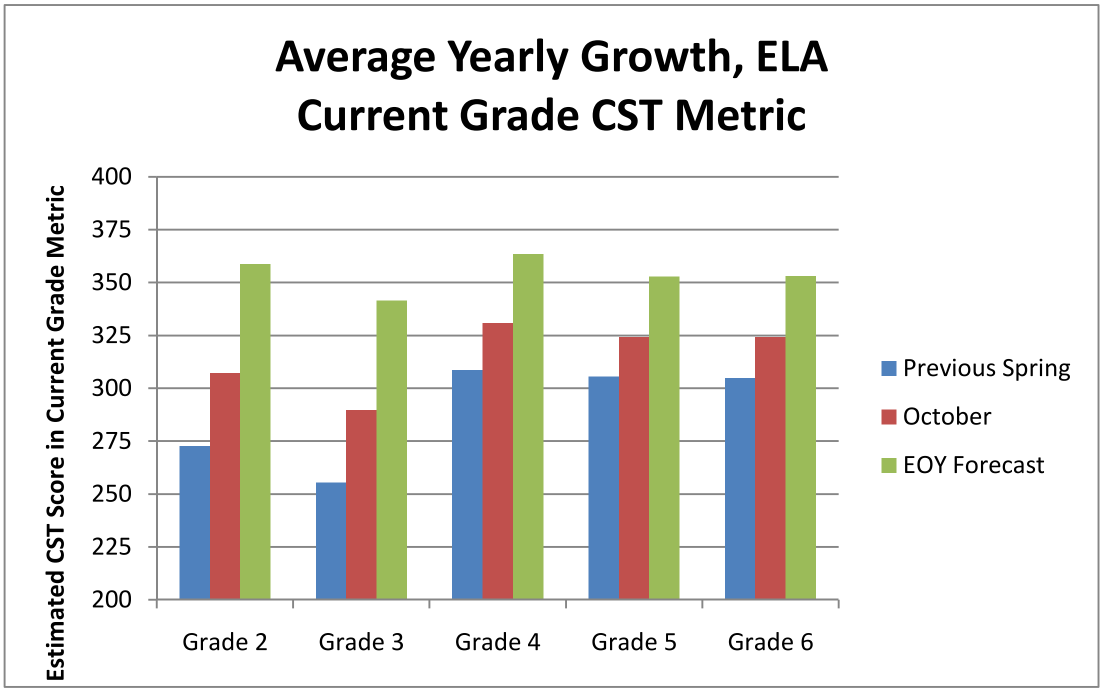
For the sake of simplicity, not all of these scores are automatically reported in STARS. However, together they make it possible with one benchmark exam to assess where the student was last year, where she is now, and where we expect her to be at the end of this year, all on a common metric that allows measurement of growth. Figures 5 and 6 illustrate what this looked like for District A.

Figure 5: Mean CST-Equivalent Scale Scores, May 2008 – May 2009

**Cross-Year Growth, CST Metric of Current Grade**

| Grade | Previous Spring | October | EOY Forecast | Total Yearly Growth |
|---|---|---|---|---|
| 2 | 273 | 307 | 359 | 86 |
| 3 | 255 | 290 | 341 | 86 |
| 4 | 309 | 331 | 364 | 55 |
| 5 | 306 | 324 | 353 | 47 |
| 6 | 305 | 324 | 353 | 48 |
| All | 290 | 316 | 353 | 63 |

Figure 6: Graphic of Mean CST-Equivalent Scale scores, May 2008 - May 2009



The same statistics are computed for every benchmark exam. The end-of-year forecast tends to be similar from benchmark to benchmark, increasing in individual student accuracy as the year progresses and students converge on their final positions in the end-of-year student ability distribution. As successive benchmarks are analyzed, their "Current CST-Equivalents" are added to the trend-line. Figures 5 and 6 show how grades 2 and 3 have much higher growth rates than the later grades. This pattern continues through high school, where growth is flat.

As mentioned, every CST-equivalent scale score is accompanied by its corresponding performance level as published by the CDE on its website. This removes the need for districts to set proficiency cut-points. Indeed, such district cut-points become confusing distractions because they bear no consistent or defensible relation to the state's cut-points.

To assess the degree to which the EdScale performance levels match those of the state, Figures 7 – 12 cross-tabulate their frequencies. The performance levels on the left margin correspond to the October 2008 CST-Equivalent end-of-year forecasts. The performance levels on the top correspond to those that students actually received from the state based on the May 2009 administration of the language CST.

Figure 7: Cross-Tabulation, Grade 2 EOY Forecast vs. Actuals, N Persons Scoring at each Performance Level

| | | 2009 EOY CST Performance Level | | | | |
|---|---|---|---|---|---|---|
| | Grade 2 | FBB | BB | Basic | Proficient | Advanced |
| | FBB | 16 | 11 | 8 | 2 | 0 |
| **October 2008** | BB | 23 | 21 | 20 | 4 | 0 |
| **EOY Forecast** | Basic | 13 | 28 | 85 | 60 | 10 |
| | Proficient | 3 | 9 | 53 | 126 | 51 |
| | Advanced | 0 | 0 | 2 | 52 | 111 |

Figure 8: Cross-Tabulation, Grade 3 EOY Forecast vs. Actuals, N Persons Scoring at each Performance Level

| | | 2009 EOY CST Performance Level | | | | |
|---|---|---|---|---|---|---|
| | Grade 3 | FBB | BB | Basic | Proficient | Advanced |
| | FBB | 50 | 26 | 5 | 0 | 1 |
| **October 2008** | BB | 49 | 75 | 44 | 7 | 0 |
| **EOY Forecast** | Basic | 20 | 85 | 204 | 98 | 12 |
| | Proficient | 2 | 6 | 55 | 213 | 100 |
| | Advanced | 0 | 0 | 5 | 44 | 139 |

Figure 9: Cross-Tabulation, Grade 4 EOY Forecast vs. Actuals, N Persons Scoring at each Performance Level

| | | 2009 EOY CST Performance Level | | | | |
|---|---|---|---|---|---|---|
| | Grade 4 | FBB | BB | Basic | Proficient | Advanced |
| | FBB | 17 | 10 | 8 | 4 | 0 |
| **October 2008** | BB | 27 | 37 | 40 | 3 | 0 |
| **EOY Forecast** | Basic | 9 | 29 | 152 | 74 | 16 |
| | Proficient | 1 | 7 | 53 | 148 | 88 |
| | Advanced | 0 | 3 | 4 | 58 | 311 |

Figure 10: Cross-Tabulation, Grade 5 EOY Forecast vs. Actuals, N Persons Scoring at each Performance Level

| | | 2009 EOY CST Performance Level | | | | |
|---|---|---|---|---|---|---|
| | Grade 5 | FBB | BB | Basic | Proficient | Advanced |
| | FBB | 29 | 14 | 10 | 0 | 1 |
| **October 2008** | BB | 31 | 21 | 37 | 4 | 1 |
| **EOY Forecast** | Basic | 20 | 62 | 203 | 103 | 3 |
| | Proficient | 2 | 5 | 72 | 223 | 102 |
| | Advanced | 0 | 0 | 2 | 39 | 190 |

Figure 11: Cross-Tabulation, Grade 6 EOY Forecast vs. Actuals, N Persons Scoring at each Performance Level

| | | 2009 EOY CST Performance Level | | | | |
|---|---|---|---|---|---|---|
| | Grade 6 | FBB | BB | Basic | Proficient | Advanced |
| | FBB | 19 | 20 | 14 | 0 | 0 |
| **October 2008** | BB | 13 | 32 | 41 | 3 | 0 |
| **EOY Forecast** | Basic | 18 | 51 | 219 | 113 | 9 |
| | Proficient | 3 | 6 | 71 | 242 | 100 |
| | Advanced | 0 | 0 | 1 | 59 | 197 |

Figure 12: Cross-Tabulation, All Grades, EOY Forecast vs. Actuals, N Persons Scoring at each Performance Level

| | | 2009 EOY CST Performance Level | | | | |
|---|---|---|---|---|---|---|
| | All | FBB | BB | Basic | Proficient | Advanced |
| | FBB | 131 | 81 | 45 | 6 | 2 |
| **October 2008** | BB | 143 | 186 | 182 | 21 | 1 |
| **EOY Forecast** | Basic | 80 | 255 | 863 | 448 | 50 |
| | Proficient | 11 | 33 | 304 | 952 | 441 |
| | Advanced | 0 | 3 | 14 | 252 | 948 |

Figures 7 – 12 show that there is reasonable agreement between the performance levels forecasted by EdScale in October 2008 and the CST performance levels assigned by the state based on the May 2009 administration of the CST for English Language Arts. The relationship is slightly weak in the bottom performance levels, but strengthens considerably in the top three performance levels.

Figures 13 - 15 show the relationship more explicitly. Figure 13 compares the mean CST-Equivalent scale and score and standard deviation of the October forecast with those of the actual CST for each grade. Figure 14 explores their correlation and classification accuracy. Figure 15 graphs them.

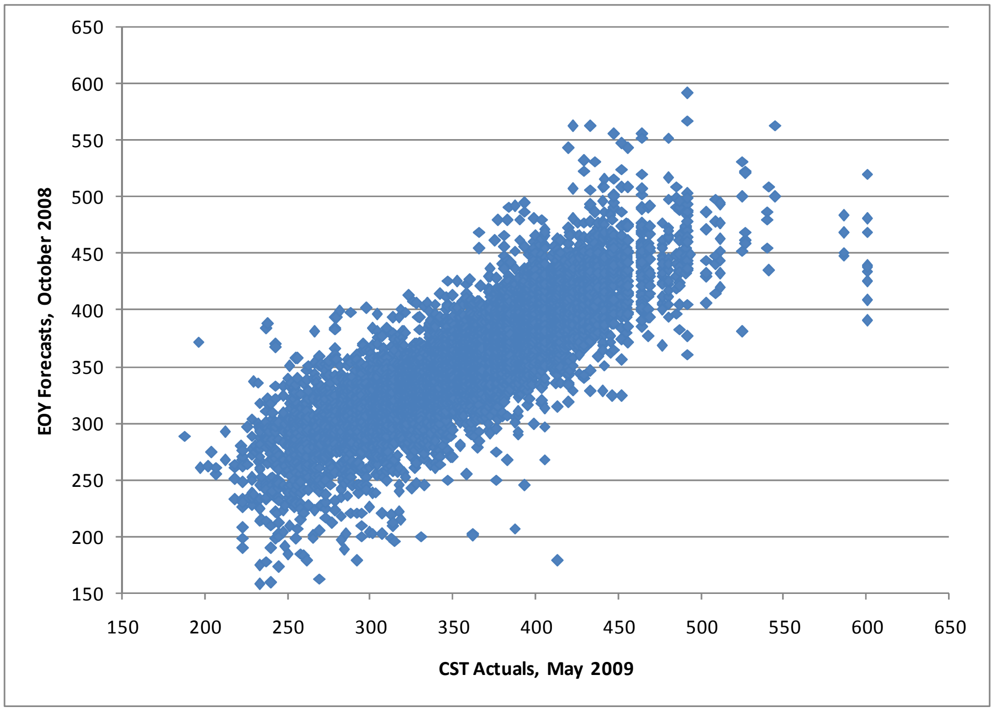
Figure 13: Forecast vs. Actuals, Mean CST-Equivalent Scale Score and Standard Deviations

| | October 2008 EOY Forecast | | 2009 EOY CST | |
|---|---|---|---|---|
| Grade | Mean SS | SD | Mean SS | SD |
| 2 | 359 | 63 | 360 | 64 |
| 3 | 341 | 61 | 347 | 63 |
| 4 | 364 | 58 | 371 | 62 |
| 5 | 353 | 53 | 359 | 57 |
| 6 | 353 | 53 | 360 | 52 |
| All | 353 | 58 | 359 | 60 |

Figure 14: Strength of Relationship: Correlation and Classification Accuracy

| Grade | Correl. | Root Mean Square Discrepancy | % Correct Classification | % Correct or Adjacent |
|---|---|---|---|---|
| 2 | 0.78 | 41.8 | 51% | 93% |
| 3 | 0.82 | 36.6 | 55% | 95% |
| 4 | 0.82 | 36.2 | 61% | 95% |
| 5 | 0.82 | 33.5 | 57% | 96% |
| 6 | 0.81 | 31.8 | 58% | 96% |
| All | 0.81 | 35.6 | 56% | 95% |

Figure 15: Grades 2 - 6 Scale Scores: Forecasts vs. Actuals, r = 0.81



The correlation between the forecast scale scores and the actual scale scores is around 0.81, typical of benchmark tests. The correlation tends to grow a few percentage points as the school year progresses. The root mean square discrepancy between the predicted end-of year scores and the actual scores is 35.6, about 70% of a performance level category. 56% of the students were correctly classified, meaning that the performance level of the forecast exactly matched that of the final test. The percentage goes up to 95% if we include adjoining performance level categories. These statistics are roughly in the range one would expect given the measurement error on both tests and the variation caused by differing student growth rates across the school year and differences in content coverage. As a point of reference, the scores of two equated English Language Arts CSTs administered to the same students on the same day would yield approximately a 0.91 correlation, compared to 0.81 for the October forecast.

## Discussion and Conclusion

EdScale has shown itself to be a useful technique for converting any formative exam or benchmark test into a proxy for the California Standards Test. This is subject to the proviso that the test is of reasonable quality and has content similar to that of either the Math or Language CSTs. In a strict mathematical sense, benchmark tests are permitted to capture multiple content areas (within Math or ELA) so long as: a) all tests being equated have the same content areas, albeit in varying proportions; b) the CST contains the same content areas or more content areas. In practice, EdScale appears to be fairly robust to violations of these assumptions due to the extensive cross-correlations between various subject matters.

EdScale's predictive efficacy is on par with that of other CST-prediction designs. Where it distinguishes itself is its ability to:

- Measure growth through the year and across grades
- Align with the CST-metric, which teachers understand
- Report performance levels automatically aligned to the state (no need to set cut-points)
- Report reliable student diagnostic statistics (not addressed here)
- Report useful item quality and targeting statistics (not addressed here)
- Return scale scores immediately (no waiting period to gather data)
- Scale existing benchmark tests (no need to purchase special tests)
- Scale tests administered in the past
- Scale complex, multidimensional tests

- Accommodate modifications from year to year, including completely new tests
- Avoid the need for an equating study
- Avoid the need for pre-tests, or post-tests
- Avoid the need for common items or persons across tests

In short, EdScale's innovation lies in the fact that it can do a great deal with very little.  Districts supply raw response-level benchmark test data with an administration date and answer key, as well as the district CST file for all grades for the previous year.   Additional specifications (grade, test type, testing population, administration date) are entered into STARS by the test administrator. Given EdScale's relatively modest data requirements and its ability to place tests on a common scale using the methods described in this paper, it is believed that it offers a practical and reliable method for school districts to draw increased benefit from their benchmark testing programs, and to meet the twin goals of formative and summative assessment using one set of instruments.

## References

Kingsbury, G. (2005). *Benchmarks and growth and success ... oh, my!* ACER: paper presentation.

Lewis, D. (2009). *Growth models for informative benchmark assessment.* CERA 2009: CERA presentation.

Moulton, M. H. (2010). *Object-oriented statistics: properties of Silsdorf's decomposition of incomplete matrices.* Morgan Hill, CA: Educational Data Systems.

NWEA paper. *Linking MAP to State Tests: Proficiency Cut-Score Estimation Procedures.* Resource at www.nwea.org/our-research/state-standards

Reckase, M. D. (2009). *Multidimensional Item Response Theory.* New York: Springer.

Wikipedia contributors (2010). *Least Squares.* Article at http://en.wikipedia.org/wiki/Least_squares.

Wiliam, D. & Leahy, S. (2009). *From teachers to schools: scaling up professional development for formative assessment.* Presented at AERA 2009.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.