The Gains from Vertical Scaling

Derek C. Briggs
Ben Domingue

July 7, 2012

Acknowledgements: The research in this paper was supported by a grant from the Carnegie Corporation.
Abstract

It is often assumed that a vertical scale is necessary when value-added models depend upon the gain scores of students across two or more points in time. This paper examines the conditions under which the vertical scale process would be expected to have a significant impact on normative interpretations using gain scores. It is shown that a significant change in the ordering of schools will only be observed when there is a large degree of scale shrinkage across grades. Because the magnitude of scale shrinkage necessary to produce this impact is rarely observed in practice, in most cases the approach taken to create a vertical scale is unlikely to have an impact on the ordering of school or teacher estimates of value-added. Because the purpose of vertical scaling is to facilitate inferences about growth in absolute magnitudes, the presence of a vertical scale is most relevant when questions are being posed about the magnitudes of student-level growth trajectories in some absolute sense.
The Gains from Vertical Scaling

Introduction

The key input for any value-added model is longitudinal data from standardized assessments that have been administered to students over two or more points in time. Because psychometricians often go through considerable effort to link test scores to facilitate score comparability across grades (i.e., vertical scaling) and because there are many different ways to go about this process (Tong & Kolen, 2007; Author, 2009), it is intuitive to assume that vertical scaling can have an impact on inferences about value-added. This would seem to especially be the case when a value-added model specifies a gain score as the outcome of interest. For example, according to Ballou, Sanders & Wright (2004):

Measuring student progress requires controlling in some fashion for initial level of achievement. This is done most transparently if the pre- and post-tests are on the same achievement scale (“vertically equated”), in which case the analysis can be based on simple differences or gain scores… The TVAAS [Tennessee Value Added Assessment System] requires tests that are vertically linked—scores for fourth graders, for example, must be expressed on the same developmental scale as scores for third graders, fifth graders, etc. In order to compare the progress of students over time, test forms must be equated across years. (p. 38, 43).

Similarly, McCaffrey et al. (2003) suggest that “estimated teacher effects could be very
sensitive to changes in scaling or other alterations to test construction and vertical linking
of different test forms.” Ballou (2009) investigates the tacit assumption that tests
vertically scaled using item response theory methods have equal-interval properties, and
comes to rather pessimistic conclusions, ultimately arguing in favor of value-added
modeling approaches that would only require test scores with ordinal properties. All of
this would seem to underscore the need for careful attention to be given to the vertical
scaling of test scores when this has preceded the use of these scores as the key
longitudinal inputs in a value-added model\(^1\).

The purpose of this paper is to examine the conditions that would need to be met
before the vertical scaling process can be expected to have a significant impact on the
ordering of schools or teachers with respect to estimates of value-added. Contrary to
what might appear to be conventional wisdom based on the quotes presented above, we
show that for most practical applications of value-added modeling—even when the
outcome of interest consists of test score gains rather than test score levels—the vertical
scaling process can be expected to have minimal impact. The reason for this is that
value-added models are fundamentally normative in nature. Because of this, the ordering
of schools or teachers by gain scores is only sensitive to a substantial decrease in the
variance of student score distributions from one grade to the next (i.e., scale shrinkage;
Hoover, 1984a; 1984b; Clemans, 1993; Yen, 1986; 1988; Camilli, Yamamoto, & Wang,
1993). We establish this through theoretical argument in the first section of the paper,
and then demonstrate it empirically in the second section using longitudinal item

\(^1\)Indeed, this seems to have been one of the motivations for the value-added model extension described by
Mariano et al. (2010) entitled “A Model for Teacher Effects From Longitudinal Data Without Assuming
Vertical Scaling.” Likewise, Betebenner (2009) has argued that an advantage of his student growth
percentile methodology is that is does not require a vertical scale, or, for that matter, scale score with
interval properties.
response data with student achievement linked to schools for a medium sized state from 2003 to 2006. In the third section we present a heuristic approach to evaluate the impact of departures of a scale from the interval ideal on value-added inferences. When this approach is applied to the same data set it appears that inferences about value-added are relatively insensitive to the extent of a scale’s departure from the interval ideal. This is because value-added models focus attention primarily upon the ordering of schools and teachers, not upon the magnitudes that separate the schools or teachers being ordered. In the fourth section we establish a growth modeling context where the presence or absence of a vertical scale will matter. This context occurs when direct questions are being posed about the magnitudes of student-level growth trajectories over a five year period of time. Yet even in this context, where a vertical scale is desirable to facilitate inferences about student growth, if the model is used to order school districts in terms of estimated value added, virtually identical conclusions would be reached about the effectiveness of school districts whether or not the test scores had been vertically scaled. The paper concludes with a discussion section.

The Theoretical Framework

A Brief Overview of the Vertical Scaling Process

When using Item Response Theory (IRT) based methods (the predominant approach in large-scale assessment contexts), the process of creating a vertical scale

---

2 For more a more detailed account of IRT-based methods for creating a vertical scale, see Peterson et al, 1989; Kolen & Brennan, 2004; Author, 2009.
involves at least two implicit stages. In a first stage, the raw scores for students taking grade specific test forms are transformed through the application of an item response function. This places scores onto a logit scale with an arbitrary mean and standard deviation (SD). Give the common IRT identification constraint to set the mean and SD of the logit scale for each grade-specific test to \((0,1)\), the task in establishing the vertical scale is to designate a base grade scale and then link adjacent grades to this scale. In this paper we focus on doing so through a separate, rather than a concurrent estimation approach (Hanson & Béguin, 2002; Kim & Cohen, 2008). This is accomplished by embedding common items across grades in a given year of testing, and then, leveraging the IRT property of parameter invariance, these common items can be used to estimate linear constants that link a focal grade scale to a base grade scale. The linking constants needed for the transformation can be estimated iteratively using a characteristic curve method such as the Stocking-Lord Algorithm (Stocking & Lord, 1983).

In a second stage, additional choices are typically made to transform the vertically linked scale away from the logit metric. Sometimes the transformation is mostly cosmetic, such as when a linear transformation is used to avoid displaying test score results with negative values. But in other cases, the transformation employed may be more elaborate. For example, Kolen & Brennan (2004) describe transformations that could be made to ensure that a score scale takes on a particular distributional shape, or to “stabilize” the standard error of measurement across the full range of the scale. Finally, as

---

3 Focusing on the separate estimation approach makes the theoretical argument easier to follow since there is no closed form expression for the grade-specific transformations to a scale that take place under the concurrent approach. Empirically however, the results from taking a concurrent approach to create a vertical scale have been shown to be very similar from a separate approach, so it seems likely that the same argument we build for the impact of the separate approach on gain scores would apply to the concurrent approach.
a last step in establishing the vertical scale, test developers will typically establish the smallest unit of change along the scale, round transformed scores to the nearest integer of this unit, and designate the lowest and highest obtainable scale scores (i.e., the “LOSS” and “HOSS”) for a particular grade.

A Brief Overview of Value-Added Models

In the National Research Council report *Getting Value out of Value-Added*, Braun et al. (2010) define value-added models (VAMs) as “a variety of sophisticated statistical techniques that use one or more years of prior student test scores, as well as other data, to adjust for preexisting differences among students when calculating contributions to student test performance.” (Braun et al. 2010, 1) According to Harris (2009), “the term is used to describe analyses using longitudinal student-level test score data to study the educational input-output relationship, including especially the effects of individual teachers (and schools) on student achievement.” From these definitions, two key features of VAMs are implicit. First, all VAMs use, as inputs, longitudinal data for two or more years of student test performance. Second, VAMs are motivated by a desire to isolate the impact of specific teachers or schools from other factors that contribute to a student’s test performance. It follows from this that the output from a VAM is a numeric quantity that can be used to facilitate causal inferences about teachers or schools.

Two of the most commonly applied VAMs are based upon the use of fixed and mixed effect regression approaches respectively. We briefly present each below,

---

4 For more comprehensive reviews of value-added modeling the interested reader should consult the following monographs and published articles: Braun et al., 2010; Hanushek & Rivkin, 2010; Harris, 2009; McCaffrey et al., 2003.
focusing attention on whether the model implies the need for longitudinal test scores that have been vertically scaled. Consider first the production function approach typically invoked by economists (Hanushek & Rivkin, 2010; Todd & Wolpin, 2003). Let $Y_{igs}$ represent an end of year test score on a standardized assessment for student $i$ in grade $g$, in a classroom with teacher $t$ and school $s$. The VAM is specified as

$$Y_{igs} = a_g + \beta' X_{ig} + \gamma' Z_{ig}^{(ts)} + \sum \theta_i D_{ig} + \epsilon_{igs}. \tag{1}$$

The covariates in this model are captured by $X$, which represents one or more test scores from prior grades\(^5\), and $Z_{ig}^{(ts)}$, which could represent any number of student, classroom or school-specific covariates thought to be associated with both student achievement and classroom assignment. The $D_i$'s are dummy variable indicators for teacher or schools associated with a given student and the $\theta_i$ represents a value-added parameter. The $\epsilon_{igs}$ represents a random error term that will often be decomposed into student, classroom and teacher-specific components (McCaffrey et al., 2009). By applying some constraints on the regression coefficients the production function model can also be re-expressed such that the gain score is the student-specific outcome of interest (c.f., Rothstein, 2010). It is only in the latter case that there would appear to be a need for test scores that have been vertically scaled.

The Educational Value-Added Assessment System (EVAAS; Sanders, Saxton & Horn, 1997) has the longest history as a VAM used for the purpose of educational accountability. While a detailed presentation is outside the scope of paper, a key point of differentiation between it and the production function approach presented above can be

---

\(^5\) Depending on the current year grade of the student, the number of available prior test scores in the same subject could range anywhere from 1 (if current year grade of student $i$ is 4), to 8 (if current year grade of student $i$ is 12).
seen by writing out the equation for a single test subject in parallel to Equation 1

\[ Y_{ig} = \alpha_g + \sum_{g \leq g} \theta^*_{eg} + e_{ig}. \]  

(2)

In contrast to the fixed effects specification from the production function approach, the EVAAS represents a multivariate mixed effects model. As such, teacher “effects” for a given grade are cast as random variables with a multivariate normal distribution such that

\[ \theta^*_g \sim N(0, \tau). \]

Only the main diagonal of the covariance matrix is estimated (i.e., teacher effects are assumed to be independent across grades). The student-level error term is also cast as a draw from a multivariate normal distribution with a mean of 0, but the covariance matrix is left unstructured.

The EVAAS is often referred to as the “layered model” because a student’s current grade achievement is expressed as a cumulative function of the current and previous year teachers to which a student has been exposed. For example, applying the model above to the context of univariate longitudinal data that span grades 3 through 5 results in the following system of equations:

\[
\begin{align*}
Y_{i3} &= \alpha_3 + \theta_3 + e_{i3} \\
Y_{i4} &= \alpha_4 + \theta_3 + \theta_4 + e_{i4} \\
Y_{i5} &= \alpha_4 + \theta_3 + \theta_4 + \theta_5 + e_{i5}
\end{align*}
\]

In the model above no value-added can be computed for grade 3 because \( \theta_3 \) is confounded with variability in student achievement backgrounds. In contrast, when certain assumptions hold it is possible to get an unconfounded value-added estimate for a grade 4 teacher, \( \theta_4 \). This can be seen by substituting the first equation into the second equation in the system such that \( Y_{i4} - Y_{i3} = \alpha_4 - \alpha_3 + \theta_4 + e_{i4} - e_{i3} \). Importantly, this shows that the sufficient statistic for estimates of teacher value-added under the EVAAS
are test score \textit{gains} from one grade to the next. It is for this reason that the EVAAS (and other mixed effect modeling approaches related to it) has long been presumed to require test scores that had been vertically scaled.

How Vertical Scaling Can Affect Gain Scores

Imagine two tests administered across two adjacent grades. The two tests have been separately placed onto the logit metric using IRT. Denote the two test score scales that result by \( y \) and \( x \), where \( x \) comes from grade \( g \) and \( y \) comes from grade \( g + 1 \). The two logit scales are linked by imposing the following linear transformations

\[
\begin{align*}
x' &= \alpha_0 + \alpha_1 x \\
y' &= \beta_0 + \beta_1 y.
\end{align*}
\]

(3)

It is easy to show that these linking transformations are completely inconsequential when the production function VAM (Equation 1) is being used to estimate value-added. For example, consider the simplest VAM specification that only conditions on prior grade achievement (\( x_i \)):

\[
y_i = \gamma_0 + \gamma_1 x_i + \sum_t \theta_t D_{it} + \epsilon_i.
\]

(4)

Now consider the same model after the two scales have been vertically linked using (3):

\[
\beta_0 + \beta_1 y_i = \gamma_0 + \gamma_1 (\alpha_0 + \alpha_1 x_i) + \sum_t \theta_t D_{it} + \epsilon_i.
\]

(5)

With a little algebra (5) can be rewritten as

\[
y_i = \left( \frac{\gamma_0 - \beta_0}{\beta_1} \right) + \left( \frac{\gamma_1 (\alpha_0 + \alpha_1 x_i)}{\beta_1} \right) + \sum_t \theta_t D_{it} + \frac{\epsilon_i}{\beta_1}.
\]
where $\theta_i' = \frac{\theta_i}{\beta_i}$. It follows that the value-added parameters $\theta_i$ and $\theta_i'$ from (4) and (5) will be perfectly correlated.

In contrast, consider the case where $\gamma_i = 1$ such that the outcome variable of interest is a gain score. Before vertical links have been established, we have

$$y_i - x_i = \gamma_i + \sum \theta_i D_i + \epsilon_i.$$  \hspace{1cm} (6)

Once again, consider the same model after the two scales have been vertically linked:

$$\beta_i y_i - \alpha_i x_i = \gamma_i + \sum \theta_i D_i + \epsilon_i.$$ \hspace{1cm} (7)

In this case, unless the two multiplicative linking constants on the left hand side of the equation, $\alpha_i$ and $\beta_i$, are identical, there is no guarantee that the value-added estimates from (6) will be linearly related with those from (7). By contrast, the values of the additive linking constants, $\alpha_0$ and $\beta_0$, will have a uniform impact that will leave inferences about value-added unchanged.

In other words, the impact of vertical linking on gain score interpretations comes through changes in the variability of the scale from grade to grade. If additional transformations are made in the process of establishing the final form of the scale, this might further compound the situation, but the basic message remains the same: only transformations that increase or decrease the variability of the scale across grades should be expected to have an impact on inferences related to gain scores. The key practical question is whether it is likely, for any pair of adjacent grades, that the vertical scaling process could lead to differences between equations (4) and (5) large enough to significantly change the ordering and/or classifications of teachers or schools. This is the
Empirically Observed Shifts in Grade to Grade Variability from Existing Vertical Scales

To get a better sense for the shifts in variability that are plausible after tests have been vertically scaled, one can examine the empirical differences in the SDs of score distributions across grades 3 to 8 in English Language Arts and mathematics respectively for 16 states. We accomplished this by gathering publicly available information about grade by grade scale score means and SDs for these 16 states from technical reports covering the years of 2007 and 2008 (Author, in press). For each state and test subject, all grade-specific SDs are divided by the grade 3 SD, and then differences are computed across adjacent grades. The summary statistics for grade to grade SD changes are shown in Table 1.

--Insert Table 1 here--

For the average state, the change in variability across grades is generally very small (between about .01 and .06 in absolute magnitude). The largest decrease in SDs across grades that was observed for any state was -0.29 in ELA (grades 3 to 4) and -0.18 in math (grades 6 to 7). The largest observed increase was 0.23 in ELA (grades 6 to 7) and 0.22 in math (grades 7 to 8).

To connect this back to the theoretical argument established in the previous section, recall the gain score model represented by Equation 7:
\[ \beta_i y_i - \alpha_i x_i = \gamma_0 - \alpha_0 - \beta_0 + \sum \theta_i D_i + \epsilon_i. \]  

Given two grades where the lower grade scale (x) has been fixed to have SD = 1, when the variability of the upper grade scale (y) increases relative to the lower grade, it follows that \( \beta_i > \alpha_i \) (where \( \alpha_i = 1 \) in this example). By contrast, when scale variability decreases across grades, \( \beta_i < \alpha_i \). So the largest decrease in grade to grade SDs shown in Table 1 of \(-0.29\) is akin to finding that \( \beta_i = 0.71 \) and \( \alpha_i = 1 \).

An Empirical Demonstration

Data

To examine the impact that changes in scale variability can have on the computation of school-level gains, we begin by replicating the process of creating a vertical scale using the empirical data from an existing state’s criterion-referenced large-scale assessment in reading (Author, 2009). The longitudinal item responses under consideration here were administered to students in grades 3 through 7 between 2003 and 2006. The vertical scale for this reading assessment was originally established by the state’s test contractor in 2001 on the basis of a common item nonequivalent groups linking design (Kolen & Brennan, 2004). The vertical score scale created for use in the present study derive from data that were obtained directly from the state’s department of education. There are two student cohorts of interest. The first cohort included students who were in grade 3 in 2003 and grade 6 in 2006; the second cohort included students who were in grade 4 in 2003 and grade 7 in 2006. The data from these two cohorts of
students are used to mimic the original approach taken to create this state’s vertical scale up to through the first stage of the process. That is, using these two cohorts of students and common items between adjacent grades and years, we created a vertical score scale using the combination of a Three Parameter Logistic IRT model (3PLM; Birnbaum, 1968), maximum likelihood estimation, and separate linking. In what follows we refer to this as the “observed” scale [O] because it is closest to the vertical scale that is used by the state to capture grade to grade growth. We subsequently summarize SD patterns by grade only with respect to the first cohort of students who were in grade 3 in 2003 and grade 6 in 2006. On the observed scale (expressed in the logit metric), the grade 3 through 6 SDs were 1, 0.87, 0.85 and 0.94.

Next, four new scales were created through successive grade-specific scale transformations that were applied in order to change the patterns of grade to grade variability.

1. Constant SD [C]: Mean growth from grade to grade transformed to follow a linear trajectory, SD transformed to be constant [1, 1, 1, 1].

2. Constant Increasing SD [CI]: Grade 4 through 6 SDs transformed to increase by 0.15 each year [1, 1.15, 1.30, 1.45]

3. Nonconstant Increasing SD [NCI]: Grade 4 and 5 SDs transformed to increase by 0.10 while the grade 6 SD increases by 0.30 [1, 1.1, 1.2, 1.5]

4. Nonconstant Decreasing SD [NCD]: Grade 4 through 5 SDs transformed to decrease by 0.10 each year while grade 6 decreases by 0.30 [1, 0.9, 0.8, 0.5]

The purpose of these transformations was to intentionally create empirical scenarios that varied the shifts in scale variability from grade to grade. Note that these transformations,
though seemingly difficult to rationalize, are not inconceivable as an approach that could
be taken by psychometricians to ensure that a vertical scale has “desirable” properties.
Indeed, Kolen & Brennan (2004) and Kolen (2006) go so far as to argue that in the
context of vertical scaling, “the IRT proficiency scale also can be nonlinearily
transformed to provide growth patterns that are consistent with expected growth
patterns…Suppose a test developer believes that the variability of scale scores should
increase over grades. If the variability of the IRT proficiency estimates does not increase
over grades, a nonlinear transformation of the proficiency scale could be used that leads
to increasing variability.” (Kolen, 2006, p. 178).

--Insert Table 2 and Figure 1 here--

 Grade to grade growth trends for the resulting five scales are shown numerically
in Table 2 in terms of means and SDs, and graphically in Figure 1 in effect size units. A
key thing to compare among the five scales are the relative differences in grade to grade
SD changes. So for example, when comparing the observed vertical scale to the
transformed scale with an SD increasing by a constant amount, we observe that the SD
from grade 3 to 4 decreased by 0.13 on the observed scale but increased by 0.15 for the
transformed scale—a relative difference of 0.28. This relative difference of 0.28 can be
conceptualized as the effect of choosing one vertical scale over the other as a outcome
variable in a value-added model.

To create a common frame of reference, the observed vertical scale and the three
vertical scales that result as a consequence of transformations that increase or decrease
grade to grade variability can be compared to a scale created to have linear growth and a constant SD [C]. The latter can be regarded as a convenient base scale and is represented in Figure 1 by a solid horizontal black line. The four dashed lines represent the effect growth trajectories for the other four vertical scales. The primary factor driving the varying trajectories of these lines are differences in the magnitudes of grade to grade SD shifts.

Comparing School-level Differences in Gain Scores by Scale

The vertical scaling process should only be expected to have an impact on value-added inferences for models that rely upon gain scores when there are large shifts in grade to grade standard deviations. The empirical evidence suggests that, on average, these shifts are usually very small; however, we did find some state-specific examples of grade to grade SD shifts between 0.20 and 0.30 in both negative and positive directions. We now put this theoretical argument to the test by computing school-level test score gains for each scale and for each of the three adjacent grade pairs for our 2003 to 2006 longitudinal cohort. Of interest are the subsequent correlations of the school-level gain scores for the four scales with shifting variability to a base scale with constant variability. We use the base scale with constant variability as our frame of reference because this is akin to the pattern that would be observed if no attempt were made to create a vertical scale at all.

--Insert Figure 2 here --
The results indicate that only a large degree of scale shrinkage will have a significant impact on the ordering of schools based on gain scores. Relative to the gain scores from the base scale with constant variance, there are 12 correlations of interest (3 grade pairs crossed by 4 scales). The mean of these 12 correlations is 0.93. Only two of 12 correlations are less than 0.95: the grade 5 to 6 gain scores for the two scales where the SDs increase and decrease by 0.30 respectively. When the grade 5 to 6 SD increases by 0.30, the correlation of these school-level gain scores with the school-level grade 5 to 6 gain scores from the base scale remains quite strong at \( r = 0.86 \). But when the grade 6 SD decreases by 0.30, the correlation with the base scale gain scores drops to \( r = 0.57 \).

Figure 2 provides a panel scatterplot of the grade 5 to 6 gain scores for the latter scenarios. In the case where \( r = 0.57 \) (left panel) the presence of a large degree of scale shrinkage (decreasing variability) essentially restricts the range of gain scores, making it much harder to reliably distinguish schools on this basis. In contrast, scale expansion (increasing variability, right panel;) does not lead to the same phenomenon.

Note however, that a decrease in variability as large as 0.30 SDs was only observed in practice for one out of five adjacent grade pairings for a single state (out of 16) in one test subject. Decreases in variability—to the extent that they were observed at all—were much more likely to be somewhere between \(-0.05\) and \(-0.15\), and these would not have a significant impact on the ordering of schools as a function of average gain scores. In additional analyses not shown here, we used grade 3 through 8 data from the same state’s reading assessment and examined the correlation between school-level gains under different vertical scales created from different linking constants by fixing \( \alpha_1 \).
at 1 and let $\beta_i$ vary. For values of $\beta_i$ between .9 and 1.1, our correlations between school-level estimates were 0.97. Only for values of $\beta_i$ below 0.80 did we observe correlations that dropped below 0.90.

**Departures from the Interval Scale Ideal**

Ballou (2009) has pointed out that VAMs depend on the assumption that test scores have interval scale properties, irrespective of whether the VAM expresses the outcome variable as test score gains or test score levels. With this in mind it would be hard to argue that any of the vertically linked scales presented in the previous section have equal-interval properties. The observed vertical scale that was the source for the additional transformations described above was created by applying the 3PLM to sets of grade-specific dichotomous item responses and then linking these sets using the Stocking-Lord algorithm. The theory of conjoint measurement (Luce & Tukey, 1964; Krantz et al, 1971) provides the only analytical framework that could invoked to evaluate whether the resulting scale could be said to have interval as opposed to ordinal or nominal properties. In practice, such a rationale has seldom been applied empirically, and generally hinges upon making an analogy between the Rasch Model and a specific version of the theory of conjoint measurement known as additive conjoint measurement (Brogden, 1977; Perline, Wright & Wainer, 1979; Borsboom, 2005; Borsboom & Zand Scholten, 2008, Michell, 2008a; 2008b; Kyngdon, 2011).

However, from a pragmatic perspective, one might ask how large the departures of each scale from an equal-interval ideal would need to be before they would have an
impact on inferences about school-level value-added. To quantify the degree to which a scale departs from the equal-interval ideal one could extend an approach previously employed by Hoover (1984a) and more recently by Ballou (2009). The idea is to assess, for each of the five scales that were considered above, the amount of growth that would be required for a student to maintain her position at the 10th, 25th, 50th, 75th and 90th percentiles of the normative score distribution across adjacent grades. These magnitudes are not directly comparable across scales because of the different transformations that were imposed to create each scale. Thus, to allow for such comparisons, we follow Ballou (2009) in taking, for each pair of adjacent grades and each scale, the ratio of the gains needed to maintain a position at the 25th, 50th, 75th and 90th percentiles relative to the gain needed to maintain a position at the 10th percentile.

--Insert Table 3 here--

In the ideal case of a scale with interval properties, one might anticipate that these ratios will be close to 1, as Table 3 illustrates using the canonical example of length, an attribute than can be expressed on a scale with not only equal interval, but ratio properties. According to data from the National Center for Health Statistics, the amount of growth in inches required for boys to maintain the same position in a normative height distribution is almost the same across the five percentiles shown in Table 3. Boys whose initial height is in a higher percentile at 12 months of age have to grow about the same to maintain the same relative position compared to boys whose initial height is at a lower percentile. This supports the notion that the more that a given vertical scale has ratios
departing from 1 across starting percentiles for any given grade pair, the stronger the circumstantial evidence that the scale has properties that depart from the interval ideal. The evidence is circumstantial in the sense that one cannot rule out the possibility that a scale has interval properties despite having ratios at different percentiles that are greater or less than 1. After all, if one was to discover that 12 month old boys at the 75th percentile in height tend to grow three times as fast as boys at the 25th percentile, this would still not invalidate the units of a ruler as existing on an interval scale. Yet when these sorts of values differ dramatically as a function of the starting percentile, it suggests that something may be amiss.

--Insert Table 4 here--

Table 4 reports the same ratios of gains at the 25th, 50th, 75th and 90th percentiles gains relative to the 10th percentile gain for the five vertical scales created for this study. For the observed vertical scale [O] the four ratios associated with grade 3 to 4 growth were 1.03, 1.04, 0.98 and 0.71 respectively. It follows from this that at the 90th percentile there is some evidence against an interval score interpretation—the gains required for students to maintain their position at the 90th percentile are just 71% of the gains required to maintain their position at the 10th percentile. In general, for the observed scale we see the strongest evidence for departures from an interval interpretation with grade 5-6 gains. The scale transformed to have constant growth and variability provides an interesting contrast to the observed scale. On the whole, the ratios for this scale smaller, yet here the ratios are largest for the percentiles associated with grade 3-4 gains, and smallest for the
percentiles associated with the grade 4-5 and 5-6 gains. In general, all of the versions of the vertical scales have growth patterns that would seem to indicate significant departures from the interval ideal for at least one of the three grade pairs for which gains scores have been computed. This demonstrates that vertical scale transformations can have a notable impact on the way gain magnitudes can/should be interpreted at different points along the scale.

What is less clear is whether departures from the interval ideal will have a significant impact on the relative rankings of schools as a function of average gain scores. To get a sense for this, we first compute, for all schools in our sample, the mean grade to grade score gain as a function of the five vertical scales. For each school there are a total of 5 mean gain scores for each of three grade pairs. Next we compute all 10 pairwise correlations within a grade pair across schools as a function of the underlying scale. This produces a total of 30 correlation coefficients (10 pairwise correlations within each of three grade pairs). Higher correlations represent scale pairings where the transformation of one to the other will have less impact on school rankings. In a vast majority of cases, gains across scales are very strongly correlated: 20 out of 30 have a correlation greater than 0.90 and the median correlation is 0.96. Only four scale pairings have correlations that depart from this trend and they all involve a pairing of one scale with the scale that was transformed to have a major decrease in variance from grade 5 to 6. So while a departure from the interval scale ideal may matter greatly when interpreting growth for individual students, it does not appear to have much impact on the relative ordering of schools as a function of average score gains.
When Does a Vertical Scale Matter?

Thus far we have demonstrated that different vertical scales—with different patterns of grade to grade variability and departures from the interval ideal—are unlikely to lead to significantly different orderings of schools according to their gain scores. The crux of the issue is that the purpose of vertical scaling is to facilitate inferences about growth in absolute magnitudes, while the purpose of value-added modeling is to facilitate inferences about teacher or school effectiveness in a normative sense. It follows that the act of establishing a vertical scale is only relevant when questions are being posed about the magnitudes of student-level growth trajectories. To help illustrate this, consider the following set of research questions that could be posed using the longitudinal grade 5 to 9 reading achievement data from students who attended public school districts in a medium-sized state between 2003 and 2008:

1. What was the average annual growth rate of students in reading?

2. Do growth rates differ significantly as a function of
   a. Gender?
   b. Free and Reduced Lunch eligibility status?
   c. English Language Learner status?
   d. Special Education status?
   e. Gifted and Talented status?

3. Do initially low-achieving students in grade 5 grow faster in reading than initially high-achieving students?

4. How do school districts rank with respect to the average growth of their students?
One relatively sophisticated way to address these questions would be to specify a three-level hierarchical linear model (Raudenbush & Bryk, 2002), where a linear growth function for repeated measures (reading test scores from grades 5 to 9) is nested within students who are nested within school districts. The three level model is

\[ Y_{ijk} = \pi_{0jk} + \pi_{1jk} \text{GRADE}_{ijk} + e_{ijk} \]  

(8.1)

\[ \pi_{0jk} = \beta_{00k} + \beta_{01}' \text{X}_{jk} + r_{0jk} \]
\[ \pi_{1jk} = \beta_{10k} + \beta_{11}' \text{X}_{jk} + r_{1jk} \]  

(8.2)

\[ \beta_{00k} = \gamma_{000} + \theta_{00k} \]
\[ \beta_{10k} = \gamma_{100} + \theta_{10k} \]  

(8.3)

where

\( Y_{ijk} \) is the test score in grade \( i \) for student \( j \) in school district \( k \) expressed on scale \( s \);

\( \text{GRADE}_{ijk} \) is an indicator variable for grades 5 through 9, recoded to go from 0 to 4;

\( \text{X}_{jk} \) is a vector of student-level dummy variables for gender, free and reduced lunch eligibility, English language status, special education status, and gifted and talented status;

\( \gamma_{000} \) and \( \gamma_{100} \) are fixed effect coefficients associated with the level 1 intercept \( \pi_{0jk} \) and slope \( \pi_{1jk} \);  

\( \beta_{01} \) and \( \beta_{11}' \) are vectors of fixed effect coefficients that interact with the level 1 intercept \( \pi_{0jk} \) and slope \( \pi_{1jk} \);

\( e_{ijk} \) represents a random grade-specific deviation for student \( j \) in district \( k \) from the conditional mean;
$r_{0,jk}$ and $r_{1,jk}$ represent level 2 random effects (student-specific deviations from the conditional means for the intercept and slope parameters); and

$\theta_{00k}$ and $\theta_{10k}$ represent level 3 random effects (district-specific deviations from the conditional means for the intercept and slope parameters).

In a value-added modeling context, the parameter $\theta_{10k}$ would typically be given the interpretation as the school district effect on student achievement, as it represents an increment in reading achievement growth that is either above or below the average for the entire state. The random effects in the model above are assumed to be independent across levels and drawn from either a univariate or multivariate normal distribution with an unstructured covariance matrix.

To simplify the illustration, only students who remain in the same school district from grades 5 to 9 and who were tested in each grade are included in the analysis. In addition, student-level covariates are fixed to take on the whatever value was observed for a given student as of grade 5. This leaves us with a sample of 20,062 students from 174 distinct school districts. Of these students, 54% were female, 25% were eligible for free or reduced lunch services, 5% are classified with limited English proficiency, 1% with no English proficiency, 8% receive special education services, and 13% were identified as gifted and talented.

We estimate the parameters from the model above using the R package lme4 (Bates, Maechler & Bolke, 2012) with three different versions of the longitudinal test score outcome. In version 1 ("z score"), we sum together the number of multiple choice items a student has answered correctly in a given grade, and then standardize the resulting variable. As a result the z score outcome variable has a mean of 0 and an SD of
1 across grades 5 through 9. In version 2 (“theta”), we transform the response pattern for each student in a given grade to an estimate of ability using the IRT 3PLM with maximum likelihood estimation. As a result the theta outcome variable has a mean of about 0.25 logits and an SD of about 1 across grades 5 through 9. Finally, in version 3 (“vertical scale”) we take the ability estimates from version 2 and link them together across grades to create a vertical scale (i.e., thereby recreating the “observed” scale from the previous section, but this time with 5 as the base grade). The grade 5 through 9 means for this scale, in logits, are 0.21, 0.67, 1.17, 1.55, 1.88 and the SDs are 0.95, 0.98, 0.94, 0.91, 0.79.

Clearly, the concept of growth is entirely different for the z score and theta scales relative to the vertically linked scale. For the first two scales growth is purely normative—a student with higher scores from one grade to the next is a student whose achievement has improved over time relative to her peers. As such for these scales it is difficult to make meaningful statements about the average “rate” of growth—by definition, this growth rate is 0. By contrast, according to the vertical scale the growth rate of the average student is about 0.42 logits per grade, which represents about 44% of the grade 5 SD and 53% of the grade 9 SD.

--Insert Table 5 here--

The HLM parameter estimates for each scale are presented in Table 5. The fixed

---

6 The mean is slightly greater than 0 in this case because the 3PLM was initially applied to the full population of students in the state before the restriction was made to limit the analysis to the subsample of 20,062. This implies that students who left school districts during this time frame tended to be slightly lower achieving than those who stayed in the same school district throughout.
effects under the row heading “Grade 5 Achievement” can be interpreted as the average grade 5 achievement as a function of student level covariates. The fixed effect for slope under the row heading “Annual Growth Rate from Grade 5 to 9” represents the average annual growth rate as a function of student level covariates. The reference categories for the fixed effects are female students in the state who are not eligible for free and reduced lunch services, are native English speakers, and not classified as either gifted and talented or receiving special education services. The 7 main fixed effect coefficients associated with grade 5 achievement levels are almost identical regardless of scale because they all reference student performance across districts in grade 5. Where the interpretation of fixed effect coefficients varies is when they are interacted with growth rates (under the row heading “Annual Growth Rate from Grade 5 to 9”). Here we see that inferences about average differences in growth as a function of student characteristics can change significantly when the frame of reference of a scale shifts from normative to absolute. For example, consider students who were classified as gifted and talented (GT) in grade 5. On the basis of the z score scale, the average GT student grows by an additional 0.07 SDs in reading achievement each grade, so cumulatively from grade 5 to 9 she will have gained an additional 0.28 SDs (i.e., 4*0.07 = 0.28) relative to her peers. On the basis of the theta scale, the achievement of the average GT student stays about the same from grade to grade relative to her peers (the model predicts a cumulative marginal decrease from grade 5 to 9 of .04 SDs). But compared to her non-GT peers on the basis of the vertical scale, the average GT student grows at a significantly slower rate in an absolute sense. By grade 9 a GT student is predicted to have grown about 0.12 logits less than a non-GT student, which is about 15% of the grade 9 SD. As this example demonstrates,
the creation of a vertical scale will have a substantive impact when comparisons of
growth are desirable on the basis of absolute magnitudes. For a different example,
consider students receiving special education services. According to the z score scale
these students are showing dramatic growth relative to their peers—cumulatively the
average student receiving special education services grows almost half of a grade 9 SD
more than students who are not receiving special education services. However, this
marginal increase in growth appears much less impressive on the theta scale and on the
vertical scale. According to the vertical scale these students only grow about 0.20 logits
more than students not receiving special education services from grades 5 to 9, which
represents about 25% of a grade 9 SD. This is still notable, but only half as large in
magnitude relative to the results implied by the z score scale.

Note that even in a normative sense, growth is consistently smaller on the theta
scale than it is on the z score scale. One plausible explanation for this is that many
student subgroups who are significantly below average in achievement as of grade 5 are
much more likely to not only guess on the multiple-choice items given to them on their
reading assessment, but to become better at guessing the correct answers over time (one
cause of this might be teachers coaching students on how to take standardized tests). The
use of the 3PLM to scale response patterns may adjust for this spurious source of growth.

Do low achieving students in grade 5 grow faster than high achieving students?
For the two normative scales, the answer to this is “not really”: the correlation between
the student-level intercept and slope is −0.27 for the z score scale and −0.19 for the
theta scale. The answer is different for the vertical scale, where the respective correlation
is −0.51, indicating that on average, lower achieving students in grade 5 grow more
through grade 9 than higher achieving students.

Finally, what about value-added inferences? For each district we can retrieve Empirical Bayes estimates of the random effect $\theta_{10k}$. The intercorrelations among the estimates across the three scales are

- 0.91 for the z score scale and theta scale,
- 0.85 for the z score scale and vertical scale, and
- 0.87 for the theta scale and vertical scale.

In this example, one would reach largely the same conclusion about the ordering of school districts even if standardized raw scores were used in place of scale scores that had been calibrated or calibrated and linked using IRT methods. In other words, the presence or absence of a vertical scale does not significantly change inferences about value-added.

Discussion

The purpose of value-added models is to support inferences about the effects of teachers and/or schools on student achievement. But these effects have a fundamentally normative interpretation—a school is considered “effective” if the value it appears to have added to student achievement is significantly larger than the average for all other schools to which it is being compared. Because of this, additive changes to a test score scale from grade to grade will not have an impact on value-added inferences. This is true even when a value-added model uses gains scores as a dependent variable; the ordering of teachers and schools as a function of average gain scores is only sensitive to scale
transformations that lead to significant decreases in score variability across grades. It follows from this that the decision to create a vertically linked score scale will only have an impact on value-added inferences based on gain scores when the process leads to substantial scale shrinkage relative to what would have been observed if a different approach had been taken to create the vertical scale, or if the scores not been linked at all. This was shown to be the case theoretically and the demonstrated empirically. When school-level gain scores from grade 5 to 6 were computed for two vertical scales—one that had been transformed to have constant variability across grades and the other transformed to have a 0.30 SD decrease—there was a significant change in the ordering of schools from one scale to the other.

This comparison captures a worst case scenario if scale shrinkage represented the empirical truth about student achievement over time. Suppose for a sequence of tests across grades that if a vertical scale were to be established, one would in fact observe substantial scale shrinkage. Suppose that instead grade-specific test scores are (a) standardized within each grade, or (b) calibrated using an IRT model but not linked. In case (a), the variability of scores across grades would stay constant by definition; in case (b), because of the typical IRT N(0,1) identification constraint on the population distribution of ability it would also stay roughly constant. In either case, true scale shrinkage would be obscured by not creating a vertical scale, and this would distort inferences about value-added.

If the process of establishing a vertical score scale could always be trusted to provide test users with insights about the empirical reality of scale score variability, then it would always be prudent to create a vertical scale to underlie the computation of gain
scores. A recent review of existing vertical scales examined 160 different grade to grade SD changes (16 states * 5 grade pairs * 2 test subjects) and found only one example of scale shrinkage strong enough to have a significant impact on gain score orderings. Nonetheless, such an occurrence, while rare, is still possible. A potential problem with this line of reasoning is the notion that it is admissible for vertical scales to be monotonically transformed to ensure that grade to grade variability stay constant or increase (Kolen & Brennan, 2004). If vertical scales are manipulated in this manner by test developers, than it implies the test scores have only ordinal properties. This is rather peculiar as it would seem inconsistent with the entire purpose of creating a vertical scale to make it possible to compare students in terms of changes in magnitude (Author, 2010).

To a large extent, the issue of whether a scale can be treated as though it has interval properties is prior to the issue of whether or not scales for adjacent grades can be linked together. Along these lines Ballou (2009) has argued that departures from the interval scale ideal could create serious problems for any of the commonly used value-added models presented in section 2 of this paper, because most of them make the implicit assumption that the outcome variable is a continuous, equal-interval variable. There are, in fact, rigorous ways that such an assumption could be tested. This has been discussed in some detail by Author (in press), and advances to the methodology have been developed by Author (2012). In the present paper we presented a less rigorous but much more easily implemented heuristic that can be used to establish the extent to which a given scale departs from the interval ideal. The basic idea is to compare competing scales with respect to departures from a percentile gain ratio of 1. The empirical question is whether observing a scale with a greater departure from the ideal has an impact in any
pragmatic sense. In the example considered here, differences in a scale’s departure from the interval ideal did not appear to have a significant impact on the ordering of schools as a function of gain scores.

Vertical scales are desirable when direct inferences are desired about how much a student has learned over two or more points in time. In this paper we provided the example of specifying a linear growth curve model with three different reading outcome scales, two that were normative in nature and a third that had been vertically scaled. The choice of scale led to substantively different answers to questions such as “Do students receiving special education services grow faster in their reading achievement than students who do not receive special education services?” For the two normative scales, questions about how much the average student has grown must be reconceptualized in terms of how much the average student’s achievement has increased relative to her peers. Nonetheless, note that when the growth curve model was used to generate value-added estimates at the district level, choice of scale had virtually no impact on the ordering of districts. This reprises the theme that has been established throughout: when two or more repeated measures are being used for the primary purpose of ordering teachers, schools or (in this instance) school districts normatively, the presence or absence of a vertical scale is unlikely to make much difference.

It is possible that choices in vertical scaling would have a more significant impact when they are used as a basis for simple linear models that project student achievement into the future. For example, in some states, a vertical scale is used as a means of setting vertically articulated cut-points across grades through the process of standard-setting. Since projections of student achievement are evaluated relative to these cutpoints, if two
different vertical scales led to different cutpoint locations, this could change the cumulative distribution of students below a given cutpoint. But in general, vertical scales seem much more likely to facilitate meaningful interpretations about growth when the focus is on individual students rather than the teachers or schools in which they are situated. If careful thought is given ahead of time to the proper design and validation of vertical scales, they could play an important role in communicating growth trends at the student-level in a manner that is intuitive to stakeholders, and providing complementary information that could serve as a criterion-referenced validity check on normative value-added inferences (Author, 2011). However, what vertical scaling will not do, except in rather unlikely circumstances, is lead to significant changes in the ranking of teachers and schools as a function of value-added.
References

Author (2009)
Author (2010)
Author (2011)
Author (2012)
Author (in press)

Bates, D., Maechler, M. & Bolker. B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. http://CRAN.R-project.org/package=lme4


Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical Theories of Mental Test Scores (pp. 397-479).


Borsboom, D., & Scholten, A. Z. (2008). The Rasch model and conjoint measurement...
theory from the perspective of psychometrics. *Theory & Psychology*, 18(1), 111.


measurement, vol. 1: Additive and polynomial representations. New York:
Academic Press

Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of
test performance. *British Journal of Mathematical and Statistical Psychology*,
64(3), 478-497.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of


the Carnegie Corporation.


Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint

MacMillan.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and
Data Analysis Methods*. (J. De Leeuw, Ed.) Advanced quantitative techniques in the


Table 1.
Summary Statistics for Changes in SDs of Vertical Scales Across Adjacent Grade Pairs

<table>
<thead>
<tr>
<th></th>
<th>Grade 3-4</th>
<th>Grade 4-5</th>
<th>Grade 5-6</th>
<th>Grade 6-7</th>
<th>Grade 7-8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ELA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-0.06</td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.04</td>
</tr>
<tr>
<td>SD</td>
<td>0.09</td>
<td>0.06</td>
<td>0.07</td>
<td>0.09</td>
<td>0.10</td>
</tr>
<tr>
<td>Min</td>
<td>-0.26</td>
<td>-0.11</td>
<td>-0.13</td>
<td>-0.17</td>
<td>-0.29</td>
</tr>
<tr>
<td>Max</td>
<td>0.17</td>
<td>0.12</td>
<td>0.11</td>
<td>0.23</td>
<td>0.10</td>
</tr>
<tr>
<td><strong>Math</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-0.03</td>
<td>0.00</td>
<td>0.03</td>
<td>-0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>SD</td>
<td>0.08</td>
<td>0.07</td>
<td>0.06</td>
<td>0.07</td>
<td>0.10</td>
</tr>
<tr>
<td>Min</td>
<td>-0.15</td>
<td>-0.10</td>
<td>-0.07</td>
<td>-0.18</td>
<td>-0.13</td>
</tr>
<tr>
<td>Max</td>
<td>0.12</td>
<td>0.18</td>
<td>0.14</td>
<td>0.07</td>
<td>0.22</td>
</tr>
</tbody>
</table>

N = 16 states; ELA = “English Language Arts”
Table 2.
Descriptive Statistics for Vertical Scale Transformations

<table>
<thead>
<tr>
<th>Transformation to Scale SD</th>
<th>Statistic</th>
<th>Grade 3</th>
<th>Grade 4</th>
<th>Grade 5</th>
<th>Grade 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Scale [O]</td>
<td>Mean</td>
<td>0.063</td>
<td>0.461</td>
<td>0.739</td>
<td>0.904</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.000</td>
<td>0.868</td>
<td>0.848</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td>Growth</td>
<td>---</td>
<td>0.43</td>
<td>0.32</td>
<td>0.18</td>
</tr>
<tr>
<td>Constant [C]</td>
<td>Mean</td>
<td>0.153</td>
<td>0.456</td>
<td>0.758</td>
<td>1.061</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Growth</td>
<td>---</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>Constant Increasing [CI]</td>
<td>Mean</td>
<td>0.153</td>
<td>0.524</td>
<td>0.986</td>
<td>1.538</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.000</td>
<td>1.150</td>
<td>1.300</td>
<td>1.450</td>
</tr>
<tr>
<td></td>
<td>Growth</td>
<td>---</td>
<td>0.34</td>
<td>0.38</td>
<td>0.40</td>
</tr>
<tr>
<td>Nonconstant Increasing [NCI]</td>
<td>Mean</td>
<td>0.153</td>
<td>0.501</td>
<td>0.910</td>
<td>1.596</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.000</td>
<td>1.100</td>
<td>1.200</td>
<td>1.504</td>
</tr>
<tr>
<td></td>
<td>Growth</td>
<td>---</td>
<td>0.33</td>
<td>0.35</td>
<td>0.50</td>
</tr>
<tr>
<td>Nonconstant Decreasing [NCD]</td>
<td>Mean</td>
<td>0.153</td>
<td>0.410</td>
<td>0.607</td>
<td>0.526</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.000</td>
<td>0.900</td>
<td>0.800</td>
<td>0.496</td>
</tr>
<tr>
<td></td>
<td>Growth</td>
<td>---</td>
<td>0.27</td>
<td>0.23</td>
<td>-0.12</td>
</tr>
</tbody>
</table>

Notes: Means and SDs in logits. Growth is expressed in effect size units as upper grade mean less lower grade mean divided by average SD.
Table 3. 
*Canonical Example of a Scale with Interval Properties: Length*

<table>
<thead>
<tr>
<th>Months</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>28.35</td>
<td>29.00</td>
<td>29.75</td>
<td>30.50</td>
<td>31.25</td>
</tr>
<tr>
<td>24</td>
<td>32.60</td>
<td>33.50</td>
<td>34.50</td>
<td>35.40</td>
<td>36.25</td>
</tr>
<tr>
<td>36</td>
<td>35.90</td>
<td>36.75</td>
<td>37.75</td>
<td>38.80</td>
<td>39.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Months</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 to 24</td>
<td>4.25</td>
<td>4.50</td>
<td>4.75</td>
<td>4.90</td>
<td>5.00</td>
</tr>
<tr>
<td>24 to 36</td>
<td>3.30</td>
<td>3.25</td>
<td>3.25</td>
<td>3.40</td>
<td>3.50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Months</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 to 24</td>
<td>1.06</td>
<td>1.12</td>
<td>1.15</td>
<td>1.18</td>
</tr>
<tr>
<td>24 to 36</td>
<td>0.98</td>
<td>0.98</td>
<td>1.03</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Source: National Center for Health Statistics, 2000
Table 4.
*Departures from the Interval Ideal for Transformed Vertical Scales*

<table>
<thead>
<tr>
<th>Scale</th>
<th>Grades</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>3-4</td>
<td>1.03</td>
<td>1.04</td>
<td>0.98</td>
<td>0.71</td>
</tr>
<tr>
<td>O</td>
<td>4-5</td>
<td>0.93</td>
<td>0.88</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td>O</td>
<td>5-6</td>
<td>4.52</td>
<td>6.95</td>
<td>9.03</td>
<td>9.92</td>
</tr>
<tr>
<td>C</td>
<td>3-4</td>
<td>1.56</td>
<td>2.05</td>
<td>2.31</td>
<td>2.02</td>
</tr>
<tr>
<td>C</td>
<td>4-5</td>
<td>0.97</td>
<td>0.95</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td>C</td>
<td>5-6</td>
<td>1.11</td>
<td>1.13</td>
<td>1.14</td>
<td>1.06</td>
</tr>
<tr>
<td>CI</td>
<td>3-4</td>
<td>2.67</td>
<td>4.16</td>
<td>5.20</td>
<td>5.13</td>
</tr>
<tr>
<td>CI</td>
<td>4-5</td>
<td>1.20</td>
<td>1.39</td>
<td>1.51</td>
<td>1.70</td>
</tr>
<tr>
<td>CI</td>
<td>5-6</td>
<td>1.36</td>
<td>1.59</td>
<td>1.79</td>
<td>1.85</td>
</tr>
<tr>
<td>NCI</td>
<td>3-4</td>
<td>1.96</td>
<td>2.80</td>
<td>3.35</td>
<td>3.14</td>
</tr>
<tr>
<td>NCI</td>
<td>4-5</td>
<td>1.08</td>
<td>1.17</td>
<td>1.20</td>
<td>1.30</td>
</tr>
<tr>
<td>NCI</td>
<td>5-6</td>
<td>1.70</td>
<td>2.23</td>
<td>2.69</td>
<td>2.96</td>
</tr>
<tr>
<td>NCD</td>
<td>3-4</td>
<td>1.04</td>
<td>1.07</td>
<td>0.98</td>
<td>0.57</td>
</tr>
<tr>
<td>NCD</td>
<td>4-5</td>
<td>0.72</td>
<td>0.46</td>
<td>0.20</td>
<td>0.03</td>
</tr>
<tr>
<td>NCD</td>
<td>5-6</td>
<td>0.69</td>
<td>0.35</td>
<td>0.04</td>
<td>-0.30</td>
</tr>
</tbody>
</table>

Note: O = Observed Vertical Scale; C = Constant SD, CI = Constant Increasing SD, NCI = Nonconstant Increasing SD, NCD = Nonconstant Decreasing SD.
Table 5.  
*HLM Parameter Estimates by Scale of Reading Outcome Measure*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Z score Scale</th>
<th>Theta Scale</th>
<th>Vertical Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed Effects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Grade 5 Achievement</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.053</td>
<td>0.240</td>
<td>0.200</td>
</tr>
<tr>
<td>Male</td>
<td>0.101</td>
<td>0.140</td>
<td>0.100</td>
</tr>
<tr>
<td>Free and Reduced Lunch</td>
<td>-0.400</td>
<td>-0.400</td>
<td>-0.400</td>
</tr>
<tr>
<td>Limited English Proficiency</td>
<td>-0.605</td>
<td>-0.560</td>
<td>-0.500</td>
</tr>
<tr>
<td>Not English Proficient</td>
<td>-0.874</td>
<td>-0.830</td>
<td>-0.800</td>
</tr>
<tr>
<td>Gifted and Talented</td>
<td>0.805</td>
<td>1.090</td>
<td>1.000</td>
</tr>
<tr>
<td>Special Education</td>
<td>-1.178</td>
<td>-1.120</td>
<td>-1.000</td>
</tr>
<tr>
<td><strong>Annual Growth Rate from Grade 5 to 9</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.037</td>
<td>-0.010</td>
<td>0.400</td>
</tr>
<tr>
<td>Male</td>
<td>0.023</td>
<td>0.010</td>
<td>0.000</td>
</tr>
<tr>
<td>Free and Reduced Lunch</td>
<td>-0.002</td>
<td>-0.020</td>
<td>0.004</td>
</tr>
<tr>
<td>Limited English Proficiency</td>
<td>0.073</td>
<td>0.030</td>
<td>0.050</td>
</tr>
<tr>
<td>Not English Proficient</td>
<td>0.141</td>
<td>0.090</td>
<td>0.100</td>
</tr>
<tr>
<td>Gifted and Talented</td>
<td>0.070</td>
<td>-0.010</td>
<td>-0.030</td>
</tr>
<tr>
<td>Special Education</td>
<td>0.115</td>
<td>0.030</td>
<td>0.050</td>
</tr>
<tr>
<td><strong>Random Effects Variance Components</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SD e_{jk}$ (Level 1 residual)</td>
<td>0.386</td>
<td>0.457</td>
<td>0.304</td>
</tr>
<tr>
<td>$SD r_{0jk}$ (Level 2 Intercept)</td>
<td>0.702</td>
<td>0.748</td>
<td>0.696</td>
</tr>
<tr>
<td>$SD r_{1jk}$ (Level 2 Slope)</td>
<td>0.116</td>
<td>0.096</td>
<td>0.067</td>
</tr>
<tr>
<td>Correlation (Level 2 Intercept, Slope)</td>
<td>-0.274</td>
<td>-0.189</td>
<td>-0.509</td>
</tr>
<tr>
<td>$SD \theta_{00k}$ (Level 3 Intercept)</td>
<td>0.281</td>
<td>0.287</td>
<td>0.265</td>
</tr>
<tr>
<td>$SD \theta_{10k}$ (Level 3 Slope)</td>
<td>0.051</td>
<td>0.051</td>
<td>0.044</td>
</tr>
<tr>
<td>Correlation (Level 3 Intercept, Slope)</td>
<td>-0.476</td>
<td>-0.391</td>
<td>-0.525</td>
</tr>
</tbody>
</table>

| N students                           | 20062         | 20062       | 20062         |
| N districts                          | 174           | 174         | 174           |

Note: Standard errors and p-values are excluded because data consists of full population of students and given sample size, almost all p-values are < .001.
Figure 1. Growth in Effect Sizes Units for Transformed Vertical Scales
Figure 2. Scatterplots of Grade 5 to 6 Gain Scores by School as a Function of Scale. Left Panel Compares Scale with Constant SD (y-axis) to Scale with Decreasing SD (x-axis). Right Panel Compares Scale with Constant SD (y-axis) to Scale with Increasing SD (x-axis).