Journal of Educational and Behavioral Statistics
2013, Vol. 38, No. 6, pp. 551–576
DOI: 10.3102/1076998613508317
© 2013 AERA. http://jebs.aera.net

The Gains From Vertical Scaling

Derek C. Briggs
Ben Domingue
University of Colorado

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 article examines the conditions under which the scale transformations associated with the vertical scaling process would be expected to have a significant impact on normative interpretations using gain scores. It is shown that this will depend upon the extent to which adopting a particular vertical scaling approach leads to a large degree of scale shrinkage (decreases in score variability over time). Empirical data are used to compare school-level gain scores computed as a function of different vertical scales transformed to represent increasing, decreasing, and constant trends in score variability across grades. A pragmatic approach is also presented to assess the departure of a given vertical scale from a scale with ideal equal-interval properties. Finally, longitudinal data are used to illustrate a case when the availability of a vertical scale will be most important: when questions are being posed about the magnitudes of student-level growth trajectories.

Keywords: *vertical scaling, value-added models, growth models, gain scores, scale shrinkage*

Introduction

The key input for any value-added model (VAM) 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 (Briggs & Weeks, 2009; Tong & Kolen, 2007), 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 VAM 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

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

> 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. (pp. 38, 43)

Similarly, McCaffrey, Lockwood, and Hamilton (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 (IRT) 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. A concern over the need for a vertical scale seems to have been one of the motivations for the VAM extension described by Mariano, McCaffrey, and Lockwood (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 it does not require a vertical scale, or, for that matter, a scale score with interval properties.

The purpose of this article 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. When the outcome variable of a VAM is a test score level, the presence or absence of a vertical scale is inconsequential. On the other hand, when the outcome variable consists of test score gains, the vertical scaling process can have an impact, but only when the variability in student achievement on a proposed vertical scale decreases substantially from grade to grade relative to the pattern that would have been observed in an alternative scaling (i.e., scale shrinkage; Camilli, Yamamoto, & Wang, 1993; Clemans, 1993; Hoover, 1984a, 1984b; Yen, 1986, 1988; Yen, Burket, & Fitzpatrick, 1995). We establish this through theoretical argument in the first section of the article and then demonstrate it empirically in the second section using longitudinal item 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 possible impact of departures of a scale from the equal-interval ideal. 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 ideal interval scale. This is because VAMs 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 plays an important role. This context occurs when direct questions are being posed about the magnitudes of student-level growth trajectories over a 5-year period of time. Yet even in this context,

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

where a vertical scale is desirable to facilitate inferences about student growth, if the model is also being used to categorize school districts in terms of estimated value added, very similar conclusions may be reached about the effectiveness of school districts whether or not the test scores have been vertically scaled. The article concludes with a discussion section.

The Theoretical Framework

A Brief Overview of the Vertical Scaling Process

An underappreciated aspect of creating a vertical scale is the importance of design considerations. What is the construct to be measured and how is it believed to change over time? On what grounds should common items be selected that overlap adjacent grades? How well are the common items aligned with the curriculum and instructional received by students? Although these design issues are of critical importance (cf. Kolen & Brennan, 2004; Peterson, Kolen, & Hoover, 1989), they are outside the scope of the present article, which has more limited ambitions. In what follows, we will optimistically assume that a defensible theory about student growth and development underlies a collection of test items that have been written for the purpose of creating a vertical scale. Our focus will be on the subsequent steps that might be taken to calibrate and transform item responses after a field test has been administered, and the conditions under which this will have an impact on the use of test scores for value-added inferences.

When using IRT-based methods (the predominant approach in large-scale assessment contexts), this process involves at least two implicit stages. In a first stage, the raw scores for students taking grade-specific test forms are first transformed through the application of an item response function. This places scores onto a logit scale with an arbitrary mean and standard deviation (*SD*). Given 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 vertical links between the grades is to designate a base grade scale and then link adjacent grades to this scale. In this article, we focus on doing so through a separate, rather than a concurrent, estimation approach (Hanson & Béguin, 2002; Kim & Cohen, 1998). 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 to that from taking separate approach (Hanson & Béguin, 2002), 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 separate linking approach is implemented 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

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

scale. The linking constants needed for the transformation are 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 further 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 and Brennan (2004) describe transformations that could be made to ensure that a score scale takes on a particular distributional shape. Finally, as 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 VAM

In the 2010 National Research Council and National Academy of Education report *Getting Value out of Value-Added*, VAMs are defined 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" (p. 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" (p. 321). From these definitions, two key features of VAMs are implicit. First, all VAMs use, as inputs, longitudinal data for 2 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 is intended 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, 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_{it} represent an end of year test score on a standardized assessment for student i at time t. Let j, k, and m index unique classrooms, teachers, and schools, respectively. A general expression for a VAM is

$$Y_{ijkmt} = \mu_t + \lambda Y_{it-1} + \beta_{1t}\mathbf{X}_{it} + \beta_{2t}\mathbf{P}_{jt} + \beta_{3t}\mathbf{S}_{mt} + \beta_4\mathbf{Z}_{it} + \theta_{kt} + \varepsilon_{it}. \quad (1)$$

In the regression model above, the vector $\mathbf{X}_{it}$ represents student-specific participation in a school-based program (e.g., special education, gifted, and talented,

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

etc.), the vector $\mathbf{P}_{jt}$ represents classroom peer characteristics (e.g., the average prior year test performance of a student's peers in classroom j), the vector $\mathbf{S}_{mt}$ represents school-level variables,[1] and the vector $\mathbf{Z}$ captures time-invariant student characteristics. The parameter θ_{kt} represents the increment that teacher k adds to a student i's achievement in year t. The residual error term ε_{it} is assumed to be iid $\sim N(0, \sigma^2)$. A key point is that in the model above it is not necessary for test scores in year t to be on the same scale as test scores in year $t-1$. So long as the score scales have a linear relationship, differences in the scale would be reflected by changes to the values of the parameters μ_t and λ. On the other hand, if λ is constrained to equal 1, this results in a VAM with a gain score as the dependent variable. In such cases, it appears necessary for both Y_{it} and Y_{it-1} to be expressed on a common vertical scale.

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 this article, a key point of differentiation between it and the production function approach presented above can be seen by writing out the equation for a single test subject as

$$Y_{it} = \mu_t + \sum_{t^* \leq t} \theta_{t^*} + \varepsilon_{it}. \quad (2)$$

In Equation 2, the achievement of student i in year t is expressed as a linear function of a year-specific average (μ_t) and the cumulative impact of teachers that have been associated with a student over his or her years of schooling $(\theta_1, \theta_2, \ldots, \theta_t)$. The EVAAS is a multivariate mixed-effects model. As such, teacher "effect" parameters for a given grade are cast as random variables with a multivariate normal distribution such that $\theta_t\ \theta_{t^*} \sim N(\mathbf{0}, \tau)$. 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 three consecutive years (i.e., Grades 3 through 5) results in the following system of equations:

$$\begin{aligned} Y_{i1} &= \mu_1 + \theta_1 + \varepsilon_{i1} \\ Y_{i2} &= \mu_2 + \theta_1 + \theta_2 + \varepsilon_{i2} \\ Y_{i3} &= \mu_3 + \theta_1 + \theta_2 + \theta_3 + \varepsilon_{i3}. \end{aligned}$$

In the first of the three equations, μ_1 represents the average achievement of students in the base year grade and θ_1 represents the deviation from this average for students assigned to a given teacher. In this first equation, θ_1 combines both the effectiveness of a student's teacher and all other factors that could influence a

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

student's achievement (e.g., socioeconomic status, motivation, etc.). However, when certain assumptions hold about the use of past achievement to adjust for any systematic sorting of students to teachers (cf. Ballou et al., 2004), it becomes possible to interpret θ_2 and θ_3 as distinct estimates of teacher value added in Years 2 and 3. This can be seen by substituting the first equation into the second equation in the system such that $Y_{i2} - Y_{i1} = \mu_2 - \mu_1 + \theta_2 + \varepsilon_{i2} - \varepsilon_{i1}$. Since θ_1 cancels, differences in a student's achievement from Year 1 to Year 2 are attributed to a year-specific main effect $(\mu_2 - \mu_1)$, a teacher effect (θ_2), and a residual source of stochastic error $(\varepsilon_{i2} - \varepsilon_{i1})$. Although a more detailed presentation of the EVAAS and the assumptions required before estimates of θ_2 and θ_3 can be given a causal interpretation are outside the scope of this article, the main point here is to notice that the teacher effects on the right-hand side of the equations are identified by successive test score *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 have been vertically scaled.

How Vertical Scaling Can Affect Comparisons Based on 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 y comes from time t and x comes from $t - 1$, where the time units are defined by grade levels. The two logit scales are linked by imposing the following linear transformations

$$\begin{aligned} x' &= \alpha_0 + \alpha_1 x \\ y' &= \beta_0 + \beta_1 y. \end{aligned} \tag{3}$$

For each transformation, the intercept parameters α_0 and β_0 shift the entire score scale up or down by a constant amount while the slope parameters α_1 and β_1 expand the scale when $\beta_1 > \alpha_1$, or contract it when $\beta_1 < \alpha_1$. Put differently, α_0 and β_0 affect the location of the scale, while α_1 and β_1 affect the variance of the scale. Note that if x were designated as the base grade for the vertical scale, it would be customary to constrain α_0 to 0 and α_1 to 1. In practice, the linking constants in Equation 3 are usually estimated using the Stocking–Lord algorithm (Stocking & Lord, 1983), but in what follows we will treat them as if they were known.

It is easy to show that these linking transformations are inconsequential when the production function VAM (Equation 1) is being used to estimate value added. For example, consider the simplest specification with just a single cohort of students that only conditions on prior grade achievement, x_i. Let i index students and j index either a teacher- or school-fixed effect so we can write

$$y_i = \mu + \lambda x_i + \theta_j + \varepsilon_i. \tag{4}$$

Now, consider the same model after the two scales have been vertically linked using Equation 3:

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

$$\beta_0 + \beta_1 y_i = \mu + \lambda(\alpha_0 + \alpha_1 x_i) + \theta_j + \varepsilon_i. \quad (5)$$

With a little algebra, Equation 5 can be rewritten as

$$y_i = \left(\frac{\mu - \beta_0}{\beta_1}\right) + \left(\frac{\lambda(\alpha_0 + \alpha_1 x_i)}{\beta_1}\right) + \theta_j' + \frac{\varepsilon_i}{\beta_1},$$

where $\theta_j' = \frac{\theta_j}{\beta_1}$. It follows that the parameters θ_j and θ_j' from Equations 4 and 5 will be perfectly correlated. Because many of the highest profile applications of value-added modeling use test score levels rather than test score gains as the outcome of interest (cf. Chetty, Friedman, & Rockoff, 2011), the presence or absence of a vertically linked score scale is seldom a cause for alarm.

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

$$y_i - x_i = \mu + \theta_j + \varepsilon_i. \quad (6)$$

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

$$\beta_1 y_i - \alpha_1 x_i = \mu - \beta_0 + \alpha_0 + \theta_j + \varepsilon_i. \quad (7)$$

The values of the additive linking constants α_0 and β_0 will have a uniform impact that will leave any normative comparisons of value added based on θ_j unchanged. By contrast, unless the two multiplicative linking constants α_1 and β_1 are identical, there is no guarantee that the value-added estimates from Equation 6 will be linearly related to those from Equation 7. The impact of vertical linking on gain score interpretations comes through transformations that affect 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 expand or contract the variability of the scale across grades should be expected to have an impact on normative comparisons 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 6 and 7 large enough to significantly change the ordering of teachers or schools. This is the question to be explored in the next section.

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 *SD*s of score distributions across Grades 3–8 in English Language Arts (ELA) and mathematics, respectively, for 16 states with existing vertical scales. This was accomplished using

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

TABLE 1.
*Summary Statistics for Changes in Standard Deviations (*SDs*) of Vertical Scales Across Adjacent Grade Pairs*

	Grades 3–4	Grades 4–5	Grades 5–6	Grades 6–7	Grades 7–8
ELA					
M	−0.06	−0.02	−0.01	0.03	−0.04
SD	0.09	0.06	0.07	0.09	0.10
Minimum	−0.26	−0.11	−0.13	−0.17	−0.29
Maximum	0.17	0.12	0.11	0.23	0.10
Math					
M	−0.03	0.00	0.03	−0.02	0.03
SD	0.08	0.07	0.06	0.07	0.10
Minimum	−0.15	−0.10	−0.07	−0.18	−0.13
Maximum	0.12	0.18	0.14	0.07	0.22

Note: ELA = English Language Arts.
$N = 16$ states.

information gathered by Dadey and Briggs (2012) about grade-by-grade scale score means and *SD*s for 16 states from technical reports covering the years 2007 and 2008. For each state and test subject, all grade-specific *SD*s 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.

For the average state, the change in *SD* across grades is generally very small (between about .01 and .06 in absolute magnitude). The largest decrease in *SD*s across grades for any state was −0.29 in ELA (Grades 3–4) and −0.18 in math (Grades–7). The largest increase was 0.23 in ELA (Grades 6–7) and 0.22 in math (Grades 7–8).

To connect this back to the theoretical argument established in the previous section, recall the gain score model represented by Equation 7: $\beta_1 y_i - \alpha_1 x_i = \mu - \beta_0 + \alpha_0 + \theta_j + \varepsilon_i$. Given two grades where the lower grade scale (x) has been fixed to have $\alpha_1 = 1$, when the variability of the upper grade scale (y) increases relative to the lower grade, it follows that $\beta_1 > 1$. In contrast, when scale variability decreases across grades, $\beta_1 < 1$. So the largest decrease in grade-to-grade *SD*s shown in Table 1 of −0.29 is akin to finding that $\beta_1 = .71$.

Given Equation 7, when evaluating the impact of choosing a proposed vertical scale over some alternative scale (as will be done in the next section), the variable of interest for any two adjacent grades will be the difference in *SD* differences. For example, let the superscript "p" indicate a proposed vertical scale and the superscript "a" indicate an alternative scale. Now define the difference in *SD* differences as

$$\delta_g = (\beta_1^p - \alpha_1^p) - (\beta_1^a - \alpha_1^a). \quad (8)$$

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

The g subscript indexes the higher of two adjacent grades. Suppose that for the adjacent Grades 5 and 6 on a proposed vertical scale that $\alpha_1^p = 1$ and $\beta_1^p = .71$ (keeping with the example above), while for an alternative scale, $\alpha_1^a = 1$ and $\beta_1^a = 1.10$. It follows that $\delta_6 = -.39$. As will be demonstrated, the further δ_g gets away from 0 in absolute value (but particularly in the negative direction), the greater the impact on value-added rankings based on gain scores. Note that if the alternative to a vertical scale is to make no attempt to impose the transformations implied by Equation 3, then this will typically impose the constraint that $\alpha_1^a = \beta_1^a = 1$ (e.g., when all grade-specific test scores have been scaled to be standard normal).

An Empirical Demonstration

Data

To examine the impact that differences in grade-to-grade 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. 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 on the logit metric 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, the Grades 3 through 6 *SD*s were 1.00, 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].

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

2. Constant Increasing *SD* [CI]: Grades 4 through 6 *SD*s transformed to increase by 0.15 each year [1.00, 1.15, 1.30, 1.45].
3. Nonconstant increasing *SD* [NCI]: *SD*s of Grades 4 and 5 transformed to increase by 0.10, while the Grade 6 *SD* increases by 0.30 [1.0, 1.1, 1.2, 1.5].
4. Nonconstant decreasing *SD* [NCD]: Grade 4 through 5 *SD*s transformed to decrease by 0.10 each year, while Grade 6 *SD* decreases by 0.30 [1.0, 0.9, 0.8, 0.5].

The purpose of these transformations was to intentionally create empirical scenarios that varied the shifts in scale *SD*s from grade to grade. Note that these sorts of 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 and Brennan (2004) and Kolen (2006, p. 178) have argued that in the context of vertical scaling, "the IRT proficiency scale also can be nonlinearly 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." Hence, while it is unlikely that a vertical scale would be transformed from the O scale to the NCD scale above, it represents a useful extreme for the purpose of analytical comparisons of gain scores.

Grade-to-grade growth trends for the resulting five scales are shown numerically in Table 2 in terms of means and *SD*s, and graphically in Figure 1 in terms of effect size units. The horizontal axis in Figure 1 shows three adjacent grade pairings: Grades 3–4, Grades 4–5, and Grades 5–6. Growth for each grade pair is computed as an effect size by subtracting mean scale scores for each grade and dividing by the average *SD*.

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 the *SD* of scores from grade to grade are most easily compared to a scale created to have linear growth and a constant *SD* [C]. We use this scale with constant variability as a frame of reference because this represents the pattern that would be observed if no attempt were made to create a vertical scale at all. This is represented in Figure 1 by a solid horizontal black line. The four dashed lines represent the effect size growth trajectories for the other four vertical scales [O, CI, NCI, and NCD]. The primary factor driving the varying trajectories of these lines is the differences in the magnitudes of grade-to-grade *SD* shifts.

Comparing School-Level Differences in Gain Scores by Scale

Our theoretical argument is that the transformations associated with 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 relative differences in grade-to-grade *SD*s for two competing scales (δ_g is large). The

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

TABLE 2.
Descriptive Statistics for Vertical Scale Transformations

Transformation to Scale *SD*	Statistic	Grades 3	4	5	6
Observed scale [O]	*M*	0.063	0.461	0.739	0.904
	SD	1.000	0.868	0.848	0.939
	Growth	—	0.43	0.32	0.18
Constant [C]	*M*	0.153	0.456	0.758	1.061
	SD	1.000	1.000	1.000	1.000
	Growth	—	0.30	0.30	0.30
Constant increasing [CI]	*M*	0.153	0.524	0.986	1.538
	SD	1.000	1.150	1.300	1.450
	Growth	—	0.34	0.38	0.40
Nonconstant increasing [NCI]	*M*	0.153	0.501	0.910	1.596
	SD	1.000	1.100	1.200	1.504
	Growth	—	0.33	0.35	0.50
Nonconstant decreasing [NCD]	*M*	0.153	0.410	0.607	0.526
	SD	1.000	0.900	0.800	0.496
	Growth	—	0.27	0.23	−0.12

Note: Means and *SD*s in logits. Growth is expressed in effect size units as upper grade mean less lower grade mean divided by average *SD*.

empirical evidence suggests that, on average, grade-to-grade *SD* differences for existing vertical scales 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 2003–2006 longitudinal cohort of students. Of interest are the subsequent correlations of the school-level gain scores for the four scales where the *SD* expands or contracts from grade to grade relative to a base scale where the *SD* remains constant.

The results indicate that a large degree of scale shrinkage is needed in a proposed vertical scale to have a significant impact on the ordering of schools based on gain scores. When compared to the gain scores from a base vertical scale with constant variance across grades, there are 12 correlations of interest (four scales crossed by three grade pairs). Each of these correlations is shown in the cells of Table 3 along with the associated δ_g. In 10 of the 12 cases, the correlation of school-level gain scores is 0.95 or higher. The two exceptions occur when $\delta_6 = 0.30$ for the NCI scale and when $\delta_6 = -0.30$ for the NCD scale. When $\delta_6 = 0.30$, the correlation of school-level gain scores remains quite strong at $r = .86$. But when $\delta_6 = -0.30$, the correlation with the base scale drops to $r = .57$. Figure 2

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

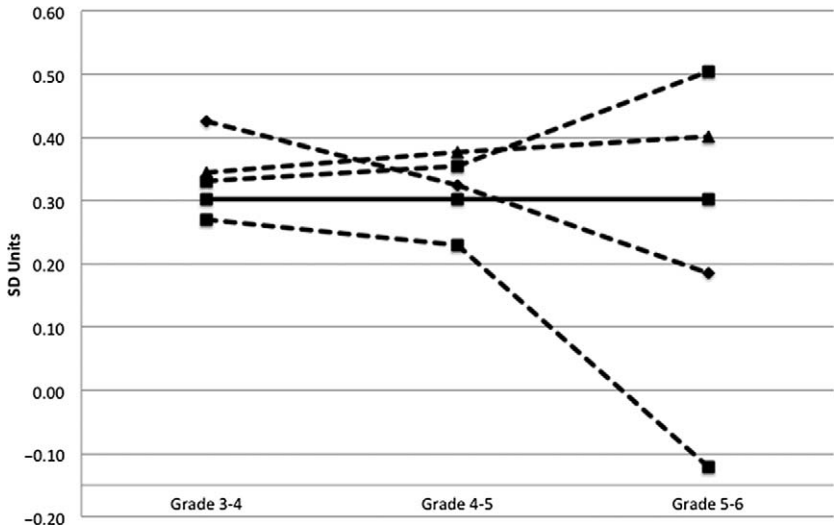

FIGURE 1. *Growth in effect sizes units for transformed vertical scales.*
Note: Effect sizes are computed for each scale and grade pair by subtracting the lower grade mean from the upper grade mean and then dividing by the average standard deviation of the two grades. The solid line represents a base scale created to show linear growth with a constant standard deviation across grades.

TABLE 3.
*The Correlations of School-Level Gain Scores as a Function of the Difference in Grade-to-Grade Standard Deviation (*SD*) Differences*

	Grades 3–4		Grades 4–5		Grades 5–6	
Proposed Vertical Scale	δ_4	*r*	δ_5	*r*	δ_6	*r*
O	−.13	.96	−.02	1.00	.09	.96
CI	.15	.96	.15	.95	.15	.96
NCI	.10	.98	.10	.97	.30	.86
NCD	−.10	.98	−.10	.96	−.30	.57

Note: The school-level gains computed for each proposed vertical scale are being compared to school-level gains for a base scale with a constant *SD* across grades. For details on the variable δ_g, see Equation 8 and accompanying narrative in text. O = observed vertical scale; C = constant *SD*, CI = constant increasing *SD*; NCI = nonconstant increasing *SD*; NCD = nonconstant decreasing *SD*.

provides a panel scatterplot of the Grades 5–6 gain scores for these two scenarios. In the case where $r = .57$ (left panel), the presence of a large degree of scale shrinkage essentially restricts the range of gain scores, making it much harder

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

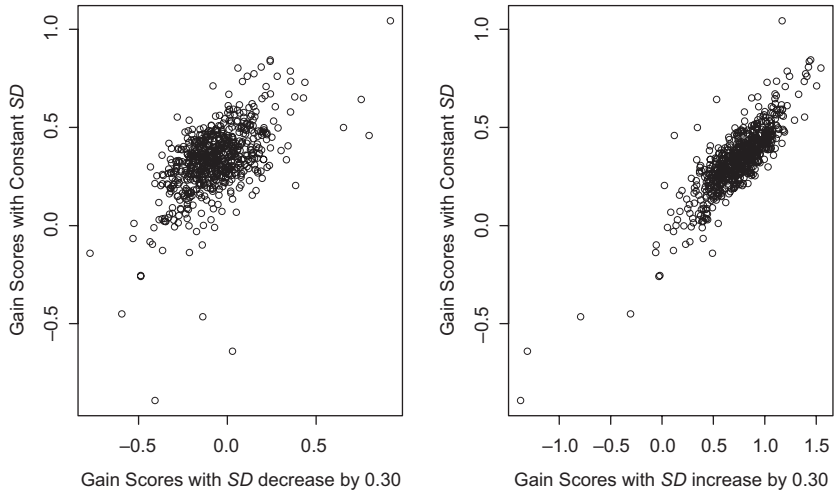

FIGURE 2. *Scatterplots of Grades 5–6 gain scores by school as a function of scale. Left panel compares scale with constant standard deviation (*SD*) (y-axis) to scale with decreasing* SD *(x-axis;* r = .57*). Right panel compares scale with constant* SD *(y-axis) to scale with increasing* SD *(x-axis;* r = .86*).*

to reliably distinguish schools. In contrast, scale expansion does not lead to the same phenomenon.

In practice, a decrease in variability as large as 0.30 *SD*s was only observed for one of the five adjacent grade pairings for a single state (of the 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 data from the same state's reading assessment across Grades 3 through 8 and examined the correlation between school-level gains under different vertical scales created from different linking constants by fixing α_1 at 1 and letting β_1 vary. For values of β_1 between 0.90 and 1.10, the correlation between school-level estimates was 0.97. Only for values of β_1 below 0.80 did we observe correlations that dropped below 0.90.

Departures From the Ideal Equal-Interval Scale

Ballou (2009) has pointed out that VAMs assume 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

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

of grade-specific dichotomous item responses and then linking these sets using the Stocking–Lord algorithm. The theory of conjoint measurement (Krantz, Luce, Suppes, & Tversky, 1971; Luce & Tukey, 1964) provides the only analytical framework that could be 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* (Borsboom, 2005; Borsboom & Scholten, 2008; Briggs, 2013; Brogden, 1977; Kyngdon, 2011; Michell, 2008a, 2008b; Perline, Wright, & Wainer, 1979).

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 an ideal equal-interval scale, one could extend an approach previously employed by Hoover (1984a, 1984b) 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 his or her position at the 10th, 25th, 50th, 75th, or 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 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, or 90th percentiles relative to the gain needed to maintain a position at the 10th percentile.

In the case of a scale with interval properties, one might anticipate that these ratios will be close to 1, as Table 4 illustrates using the canonical example of length, an attribute that can be expressed on a scale with not only 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 4. 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 3 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 may suggest some scale-dependent growth patterns that merit closer examination.

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

TABLE 4.
Canonical Example of a Scale With Interval Properties: Length

Months	10	25	50	75	90
Length of boys in inches at percentiles of national distribution					
12	28.35	29.00	29.75	30.50	31.25
24	32.60	33.50	34.50	35.40	36.25
36	35.90	36.75	37.75	38.80	39.75
Gains in inches required to stay at each percentile over a 12-month period					
12–24	4.25	4.50	4.75	4.90	5.00
24–36	3.30	3.25	3.25	3.40	3.50
Ratio of gains relative to gains at 10th percentile					
12–24		1.06	1.12	1.15	1.18
24–36		0.98	0.98	1.03	1.06

Source: Kuczmarski et al. (2002).

Table 5 reports the same ratios of gains at the 25th, 50th, 75th, and 90th percentiles 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 Grades 3–4 growth were 1.03, 1.04, 0.98, and 0.71, respectively. This implies that it is at the 90th percentile that we see the strongest 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 O scale we see the strongest evidence for departures from an interval interpretation with Grades 5–6 gains. The C scale (constant growth and variability) provides an interesting contrast to the O scale. On the whole, the ratios for this scale are smaller; yet, here the ratios are largest for the percentiles associated with Grades 3–4 gains and smallest for the percentiles associated with the Grades 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 an ideal interval scale 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 the mean grade-to-grade score gain for all schools in our sample as a function of the five vertical scales previously introduced. For each school, there are a total of five mean gain scores for each of the three grade pairs. Next, we compute all pairwise correlations *within* each grade pair as a function of the underlying scale for which the gains were computed.

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

TABLE 5.
Departures From the Interval Ideal for Transformed Vertical Scales

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

Note: O = observed vertical scale; C = constant *SD*; CI = constant increasing *SD*; NCI = nonconstant increasing *SD*; NCD = nonconstant decreasing *SD*.

This produces a total of 30 correlation coefficients (10 pairwise correlations within each of the three grade pairs). Higher correlations represent scale pairings where the transformation of one to the other will have less impact on school rankings.

We find little evidence that the rankings of schools are sensitive to departures from the ideal interval scale. That is, whether school gains are computed from a vertical scale associated with gain ratios close to 1 or with gain ratios far from 1 (see Table 5), the rankings of schools with respect to these gains remains about the same. In 20 of the 30 cases, the pairwise correlation is greater than .90 and the median correlation is .96. The pairwise correlations that most depart from this trend do not seem to be driven by apparent departures from the ideal interval scale, but by pairwise combinations of gain scores from source vertical scales with large $\delta'_g s$. For example, consider four specific pairwise comparisons of a school's Grades 5–6 gains. As a point of reference, consider the gains computed from the vertical scale that was transformed to have a 0.30 *SD* decrease from Grades 5–6 (the NCD scale). Relative to the vertical scales transformed to have either constant or increasing variance (the C, CI, NCI scales), this 0.30 *SD* decrease is associated with $\delta_6 = \{0.30, 0.45, 0.60\}$ relative to each competing scale. Not surprisingly, given the results shown in Table 3, the associated correlations in school-level gain scores decrease from 0.57, 0.31, and 0.07. The larger the relative differences in *SD* changes between two candidate vertical scales, the bigger the impact on value-added orderings.

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

When Does a Vertical Scale Matter the Most?

Thus far, we have demonstrated that different vertical scales are most likely to lead to significantly different gain score rankings when the choice of one scale over the other has a large effect on scale variability from grade to grade. 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. Since most VAMs use test score levels rather than gain scores as the outcome variable of interest anyway, the act of establishing a vertical scale will likely be most relevant when questions are being posed about the average magnitudes of student-level growth trajectories. To help illustrate this, consider the following set of research questions that could be posed using the longitudinal Grades 5–9 math 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 (GT) 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 (HLM; 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

$$Y^s_{ijk} = \pi_{0jk} + \pi_{1jk}\text{GRADE}_{ijk} + e_{ijk}. \tag{9}$$

$$\begin{aligned} \pi_{0jk} &= \beta_{00k} + \beta'_{01}\mathbf{X}_{jk} + r_{0jk} \\ \pi_{1jk} &= \beta_{10k} + \beta'_{11}\mathbf{X}_{jk} + r_{1jk}. \end{aligned} \tag{10}$$

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + \theta_{00k} \\ \beta_{10k} &= \gamma_{100} + \theta_{10k}, \end{aligned} \tag{11}$$

where Y^s_{ijk} is the test score in grade i for student j in school district k expressed on scale s; GRADE_{ijk} is an indicator variable for Grades 5 through 9, recoded to go from 0 to 4; $\mathbf{X}_{jk}$ is a vector of student-level dummy variables for gender, free and reduced lunch eligibility, English language status, special education status, and GT status; γ_{000} and γ_{100} are fixed-effect coefficients associated with the Level-1 intercept π_{0jk} and slope

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

π_{1jk}; β'_{01} and β'_{11} are vectors of fixed-effect coefficients that interact with the Level-1 intercept π_{0jk} and slope π_{1jk}; e_{ijk} represents a random grade-specific deviation for student j in district k from the conditional mean; r_{0jk} and r_{1jk} represent Level-2 random effects (student-specific deviations from the conditional means for the intercept and slope parameters); and θ_{00k} and θ_{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 θ_{10k} would typically be given the interpretation as the school district effect on student achievement, as it represents an increment in math 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 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 GT.

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 θ outcome variable has a mean[2] 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 Grades 5 through 9 means for this scale, in logits, are 0.21, 0.67, 1.17, 1.55, 1.88, and the *SD*s are 0.95, 0.98, 0.94, 0.91, 0.79.

Clearly, the concept of growth is entirely different for the z score and θ 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*.

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

The HLM parameter estimates for each scale are presented in Table 6. The fixed 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 GT or receiving special education services. The seven main fixed-effect coefficients associated with Grades 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 Grades 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 GT in Grade 5. On the basis of the *z* score scale, the average GT student grows by an additional 0.07 *SD*s in reading achievement each grade, so cumulatively from Grades 5 to 9 she will have gained an additional 0.28 *SD*s (i.e., $4 \times 0.07 = 0.28$) relative to her peers. On the basis of the θ 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 Grades 5 to 9 of .04 *SD*s). 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 θ 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 θ scale than it is on the *z* score scale. One possible 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

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

TABLE 6.
Hierarchical Linear Model (HLM) Parameter Estimates by Scale of Reading Outcome Measure

	Z Score Scale	θ Scale	Vertical Scale
Fixed effects			
Grade 5 achievement			
Intercept	0.053	0.240	0.200
Male	0.101	0.140	0.100
Free and reduced lunch	−0.400	−0.400	−0.400
Limited English proficiency	−0.605	−0.560	−0.500
Not English proficient	−0.874	−0.830	−0.800
Gifted and talented	0.805	1.090	1.000
Special education	−1.178	−1.120	−1.000
Annual growth rate from Grades 5–9			
Intercept	−0.037	−0.010	0.400
Male	0.023	0.010	0.000
Free and reduced lunch	−0.002	−0.020	0.004
Limited English proficiency	0.073	0.030	0.050
Not English proficient	0.141	0.090	0.100
Gifted and talented	0.070	−0.010	−0.030
Special education	0.115	0.030	0.050
Random effects variance components			
SD e_{ijk} (Level-1 residual)	0.386	0.457	0.304
SD r_{0jk} (Level-2 intercept)	0.702	0.748	0.696
SD r_{1jk} (Level-2 slope)	0.116	0.096	0.067
Correlation (Level-2 intercept, slope)	−0.274	−0.189	−0.509
SD θ_{00k} (Level-3 intercept)	0.281	0.287	0.265
SD θ_{10k} (Level-3 slope)	0.051	0.051	0.044
Correlation (Level-3 intercept, slope)	−0.476	−0.391	−0.525
N students	20,062	20,062	20,062
N districts	174	174	174

Note: Standard errors and *p* values are excluded because data consist of full population of students and given sample size, almost all *p* values are <.001.

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 θ 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.

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

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

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

Whether the choice of scale would have a significant impact on value-added interpretations would depend upon how these district estimates would be used. If used to rank teachers according to quintiles of the effectiveness distribution, then even a correlation as high as 0.91 could lead to significant shifts across quintiles. On the other hand, if the estimates were only to be used to categorize districts in the tails of the distribution that are significantly different from average, it is much less likely that districts would see different categorizations by choice of scale with correlations this high.

Discussion

The purpose of VAMs 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 VAM uses gain 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 had not been linked at all. This was shown to be the case theoretically and then demonstrated empirically. When school-level gain scores from Grades 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 further that instead of creating vertical links, 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

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

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 for models with gain scores as the outcome variable of interest.

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 and/or growth trajectories. A recent review of existing vertical scales examined 160 different grade-to-grade *SD* changes (16 states $\times$ 5 grade pairs $\times$ 2 test subjects) and found only two examples of scale shrinkage that would imply $\delta_g = -.20$ relative to an alternative scale with a constant *SD* across grades. Nonetheless, such an occurrence, while rare, is still possible. A potential problem with this line of reasoning is the notion that there are "true" growth trends that a vertical scale can capture. This tension becomes evident if vertical scales are monotonically transformed to ensure that grade-to-grade variability stay constant or increase. 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 facilitate comparisons of students in terms of absolute changes in magnitude.

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 an idealized interval scale could create serious problems for any of the commonly used VAMs presented in the second section of this article, because most of them make the implicit assumption that the outcome variable is a continuous variable with equal-interval properties. There are, in fact, rigorous ways that such an assumption could be tested (Briggs, 2013; Kyngdon, 2011). In the present article, 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 on intended comparison 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 to be made about how *much* a student has learned over two or more points in time. In this article, we provided the example of specifying a linear growth curve model with three different math 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 math achievement than students who do not receive special education

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

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, the choice of scale had a relatively small impact on the ordering of districts.

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 might be 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 cut points, if two different vertical scales led to different cut point locations, this could change the cumulative distribution of students below a given cut point. 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.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research in this article was supported by a grant from the Carnegie Corporation.

Notes

1. The parameters for classroom and school-level contextual variables are only identifiable when there is panel data available (i.e., multiple cohorts of students per teacher).
2. The mean is slightly greater than 0 in this case because the three parameter logistic model (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.

References

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, *4*, 351–383.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*, 37–65.

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

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

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*, 42–51.

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). Reading, MA: Addison-Wesley.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, MA: Cambridge University Press.

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

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, *50*, 204–226.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues & Practice*, *28*, 3–14.

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika*, *42*, 631–634.

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement*, *17*, 379.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (Working Paper 17699 (2), 1–51). Cambridge, MA: National Bureau of Economic Research.

Clemans, W. V. (1993). Item response theory, vertical scaling, and something's awry in the state of test mark. *Educational Assessment*, *1*, 329–347.

Dadey, N., & Briggs, D. C. (2012). A meta-analysis of growth trends from vertically scaled assessments. *Practical Assessment, Research & Evaluation*, 17. Retrieved from http://pareonline.net/getvn.asp?v=17&n=14

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy*, *4*, 319–350.

Hanson, B., & Béguin, A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*, 3–24.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*, 267–271.

Hoover, H. D. (1984a). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice*, *3*, 8–14.

Hoover, H. D. (1984b). Rejoinder to Burket. *Educational Measurement: Issues and Practice*, *3*, 16–18.

Kim, S.-H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*, 131–143.

Kolen, M. J. (2006). Scaling and norming. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 171–180). Westport, CT: American Council on Education.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer Verlag.

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, vol. 1: Additive and polynomial representations*. New York, NY: Academic Press.

Kuczmarski, R. J., Ogden, C. L., Guo, S. S., Grummer-Strawn, L. M., Flegal, K. M., Mei, Z.,... Johnson, C. L. (2002). 2000 CDC growth charts for the United States: Methods and development. National Center for Health Statistics.*Vital and Health Statistics*. *11*(246).

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

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1–27.

Mariano, L. T., McCaffrey, D. F., & Lockwood, J. R. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics*, *35*(3), 253–279.

McCaffrey, D. F., Lockwood, J. R., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability* (Vol. 158). RAND Research Report prepared for the Carnegie Corporation.

Michell, J. (2008a). Is psychometrics pathological science? *Measurement: Interdisciplinary Research & Perspective*, *6*, 7–24.

Michell, J. (2008b). Conjoint measurement and the Rasch paradox: A response to Kyngdon. *Theory & Psychology*, *18*, 119.

National Research Council and National Academy of Education (2010). *Getting value out of value-added: Report of a workshop*. Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation and Educational Accountability, H. Braun, N. Chudowsky, & J. Koenig (Eds.). Washington, DC: The National Academies Press.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*, 237.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York, NY: MacMillan.

Raudenbush, S. W., & Bryk, A. S. (2002). Advanced quantitative techniques in the social sciences. In J. De Leeuw (Ed.), *Hierarchical linear models: Applications and data analysis methods* (*Vol. 2,* pp. xxiv, 485). Newbury Park, CA: Sage.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system, a quantitative, outcomes-based approach to educational measurement. In Jason Millman (Ed.), *Grading teachers, grading schools, Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*, 3–33.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*, 227–253.

Yen, W. M. (1988). Normative growth expectations must be realistic: A response to Phillips and Clarizio. *Educational Measurement: Issues and Practice*, *7*, 16–17.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*, 299–325.

Yen, W. M., Burket, G. R., & Fitzpatrick, A. R. (1995). Response to Clemans. *Educational Assessment*, *3*, 181–190.

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015

Authors

DEREK C. BRIGGS is professor and chair of the Research and Evaluation Methodology program in the University of Colorado's School of Education, 249 UCB, Boulder, CO, 80309; e-mail: derek.briggs@colorado.edu. His research is oriented toward methods for the measurement and evaluation of student learning. To this end his research focuses on topics such as learning progressions, diagnostic assessment, item response theory, vertical scaling, growth modeling, value-added modeling and causal inference.

BEN DOMINGUE is a research associate in the Population Program within the University of Colorado's Institute of Behavioral Science, Boulder, CO, 80309; e-mail: ben.domingue@gmail.com. His research interests include value-added modeling, applied statistics and computational methods using R.

Manuscript received March 14, 2013
Revision received June 13, 2013
Accepted July 30, 2013

Downloaded from http://jebs.aera.net at ARIZONA STATE UNIV on January 9, 2015