Title: Is Growth in Student Achievement Scale Dependent?

Author(s): Derek Briggs, University of Colorado, Boulder and Damian Betebenner Center for Assessment.

Date of Initial Publication: 04/19/2009: Presented at the annual meeting of National Council for Measurement in Education, San Diego, CA, April 14, 2009.

Abstract/Summary

The impetus for this paper can be traced back to two simple questions that most parents likely ask at some point in time during their child s education: 1) How much has my child learned? 2) Is the amount my child has learned good enough? These are straightforward and intuitively important questions about *growth*. The first asks a question about the magnitude of growth. The second asks a question about criteria for judging the amount of growth. While the questions may be intuitive, the psychometric and statistical gymnastics involved in coming up with a defensible answer are not. Indeed, the deceptively simple nature of the questions masks important conceptual and philosophical undercurrents. Learning of what? How should learning be measured? Can a single number capture this phenomenon? Who decides how much learning is good enough? Once decisions about what constitutes good enough are decided, does this eliminate questions of magnitude? Or are both questions compatible, such that posing one will tend to beg an answer to the other?

In what follows we pose a research question that is also deceptively straightforward: Do interpretations of student growth depend on the way longitudinal test scores have been scaled? This paper provides a theoretical and empirical context where one can give a provisional answer of no to this question. We show that when growth interpretations are made normatively, they appear insensitive to most admissible transformation of the underlying score scale. In particular we focus attention on student growth estimates aggregated at the school-level that derive from a growth model currently being used for state accountability purposes in Colorado and Massachusetts and the value-added estimates produced by multivariate mixed-effects models. The former model relies upon quantile regression, an approach not requiring an underlying score scale with interval properties. The latter value-added model assumes interval scale properties. The two models present a contrast allowing us to examine both the within methodology impact of non-interval scales as well as the impact across methodologies.

Subject/Keywords: student achievement, student growth, score scaling, psychometrics, value-added measurement, student growth percentiles, Colorado Growth Model

Document Type: Paper

Document Archive Number: 0016cdebr2009

Is Growth in Student Achievement Scale Dependent?

Derek Briggs

University of Colorado, Boulder

Damian Betebenner

Center for Assessment

Paper presented at the invited symposium “Measuring and Evaluating Changes in Student Achievement: A Conversation about Technical and Conceptual Issues” at the annual meeting of the National Council for Measurement in Education, San Diego, CA, April 14, 2009.

We would like to acknowledge Jon Weeks, who was a big help with the programming for the analyses described in this manuscript.

Introduction

The impetus for this paper can be traced back to two simple questions that most parents likely ask at some point in time during their child's education: 1) How much has my child learned? 2) Is the amount my child has learned good enough? These are straightforward and intuitively important questions about *growth*. The first asks a question about the magnitude of growth. The second asks a question about criteria for judging the amount of growth. While the questions may be intuitive, the psychometric and statistical gymnastics involved in coming up with a defensible answer are not. Indeed, the deceptively simple nature of the questions masks important conceptual and philosophical undercurrents. Learning of what? How should learning be measured? Can a single number capture this phenomenon? Who decides how much learning is good enough? Once decisions about what constitutes "good enough" are decided, does this eliminate questions of magnitude? Or are both questions compatible, such that posing one will tend to beg an answer to the other?

Most of the psychometric and statistical obstacles to the measurement and evaluation of change in student achievement have been well-known (at least in academic circles) since the publications of Cronbach & Furby (1970), Rogosa et al., (1982), and Willet (1988) among others. What has changed with the advent of NCLB is the way that test scores and statistical models have become intermingled as a primary basis for holding teachers and schools accountable for changes in student learning (c.f., Ryan & Shepard, 2008). This intermingling has led to renewed confusions about the differences between criterion and norm-referenced interpretations of growth, and the properties of the underlying test score scales that are needed to support these interpretations.

In what follows we pose a research question that is also deceptively straightforward: Do interpretations of student growth depend on the way longitudinal test scores have been scaled? This paper provides a theoretical and empirical context where one can give a provisional answer of "no" to this question. We show that when growth interpretations are made normatively, they appear insensitive to most admissible transformation of the underlying score scale. In particular we focus attention on student growth estimates aggregated at the school-level that derive from a growth model currently being used for state accountability purposes in Colorado and Massachusetts and the value-added estimates produced by multivariate mixed-effects models. The former model relies upon quantile regression, an approach not requiring an underlying score scale with interval properties. The latter value-added model assumes interval scale properties. The two models present a contrast allowing us to examine both the within-methodology impact of non-interval scales as well as the impact across methodologies.

Background: Growth Models and Value-Added

We begin by clarifying some different meanings that are possible when one uses the umbrella term "growth model". All statistical models for test score "growth" are essentially models of conditional achievement. A key distinguishing feature is whether one wishes to model student achievement that is conditional on time, or student achievement that is conditional on prior achievement. We refer to the former as an *absolute growth model* and the latter as a *relative growth model*.

An absolute growth model can be used to answer the first of the motivating questions we posed earlier: “How much has student achievement changed from one grade to the next?” or “At what rate is student achievement changing across multiple grades?” Some examples of widely known absolute growth models are the multilevel models of change over time popularized by Raudenbush & Bryk (2002) and Singer & Willet (2004). Gain score models constitute a constrained version of such models when only two longitudinal time points are available. A key assumption of such models is that test scores have been placed onto a vertical scale to adjust for differences in difficulty such that the magnitudes of scores across grades can be directly and meaningfully compared in an absolute sense.

In contrast, relative growth models do not require a vertically linked scale, only prior test scores that are strongly associated with subsequent test scores. These models answer a different question about student achievement: “Compared to students with the same prior achievement, is current achievement higher or lower than would be expected?” This question specifies a normative answer to the first question (How much has my child learned?). The key quantities of interest in a relative growth model are residuals: the difference between any student’s observed achievement, and that which would be predicted given their prior achievement. It is this residual that provides a normative quantification of growth: growth above or below statistical expectation. The ways that these residuals are computed vary in complexity from simple linear regression models (where the conditioning of current achievement on prior achievement is most transparent) to multivariate mixed effects models (where current achievement is layered upon transformations of prior achievement), but the basic principle of relative growth as the difference between expected and observed remains the same.

As they continue to evolve, state-level systems of educational accountability are increasingly using growth models in both of the flavors described above as a means of judging the quality of teachers and schools. This has only become widely possible within the past decade in response to NCLB. Before then most states did not test students in multiple content domains annually between the elementary and high-school years. The creation of a longitudinal infrastructure through which student achievement can be linked to schools (and sometimes teachers) over time might be viewed as one positive consequence of NCLB—provided that this data is used in ways that are valid.

The leap from a growth model to what can be called a *value-added model* is a short one. It requires three steps

1. The definition of what constitutes expected test performance for a given unit of analysis (e.g., student, teacher or school).
2. The computation of some deviation from expectation that contrasts what has been observed to what would be expected for the unit of analysis.
3. The inference that the deviation from expectation reflects the "value-added" to student achievement by the unit of analysis.

While it is true that all growth models can be turned into value-added models, for relative growth models this transition is the most seamless because statistical expectation is usually implicit in such models. Indeed, for the layered model popularized by William Sanders (cf., Sanders et al., 1997), the estimation of value-added quantities is the sole purpose of the model, hence there is no transition through steps 1-3 above—the model jumps directly to step 3. For thorough discussions of the issues surrounding the specification and interpretation of value-added models see McCaffrey et al (2004), Lissitz (2005), and a forthcoming special issue of the journal *Education Finance and Policy*.

Ordinal or Interval?

When test scores are used as the longitudinal outcome measures for relative growth models, it is typically assumed that test score outcomes are continuous variables—that is, that they exist on a scale with interval properties. This is assumption is made whether or not test scores have been vertically scaled to have a developmental interpretation over time. Ballou (2008) has recently called this assumption into question in the context of using growth models to estimate the value-added effects of teachers on student achievement. At issue is whether the application of item response theory (IRT) models can and/or should be expected to produce an interval test score metric. Ballou notes—correctly we think—that there is considerable confusion in the research literature on this point.

There appear to be at least two schools of thought on this issue among psychometricians (interestingly, it is not always clear that members of each school of thought are entirely aware of the salient theoretical distinctions that may divide them). One school of thought can be represented by those who implicitly ascribe to what Michell (1986; 1990) and Hand (1996) have described as a *classical* theory of measurement. This theory might best be captured by the aphorism “If it exists, it can be measured. If it can’t be measured, it doesn’t exist.” According to Hand

> Developing a measurement procedure according to the classical theory requires relating the hypothesized quantitative attributes to observable quantities within some theoretical framework. The hypothesized quantitative attributes can then be measured by virtue of their relationships. Here the hypothesized attributes, as well as their quantitative nature, are all part of the theory being studied. Rasch’s notion of specific objectivity might be regarded as fitting naturally within this framework…According to the classical theory measurements are always real numbers: *if* (our emphasis) we have been able to measure them, the numbers which have resulted satisfy all the properties required for arithmetic manipulation, so we can manipulate them using any statistical operation. (pp. 457-458)

That is, in classical measurement there essentially is no such thing as an ordinal measure. If one is unable to quantify differences in magnitude, one is not doing measurement at all! In contrast, the second school of thought is based on the notion of *representational* measurement theory, in which empirical, qualitative data relationships are observed, and then rules are established to characterize these relationships numerically. Michell (1986) describes the core of representational theory through a quotation from Townsend & Ashby (1984):

> The fundamental thesis is that measurement is (or should be) a process of assigning numbers to objects in such a way that interesting qualitative empirical relations among the objects are reflected in the numbers themselves as well as in important properties of the number system. (p. 394)

The most famous example of this perspective is Stevens’s (1946) classification of scales as nominal, ordinal, interval and ratio. The fact that this terminology has become relatively universal in most forms of quantitative analysis is an indication that the representational view has become somewhat predominant[1]. A nod toward the representational view is also implicit in the writings of many well-known American psychometricians[2] (c.f., Harris (2007), Kolen & Brennan (2004), Thissen & Orlando (2001), Yen (1986) and Zwick (1992)).

The bridge between these two theories of measurement is the concept of conjoint measurement first established by Luce & Tukey (1964). Loosely speaking, conjoint additivity implies that two attributes can be scaled such that their additive combination forms a third measure. (The classic example used by Rasch was of the relationship between force (f), mass (m) and acceleration (a) in Newton's second law of motion after taking logarithms: A = F + M where A = log(a), F = log(f) and M = -log(m).) When conjoint additivity holds, a measured quantity (e.g.,, the log odds of a correct item response) possesses interval properties because invariant comparisons can be made between two attributes (e.g., person ability and item difficulty). This gets to the heart of a fundamental misconception rooted in the assertion that there is nothing uniquely interval about the metric of θ, which has never been the claim under conjoint measurement. As Zwick (1992) points out, it is not θ that is (necessarily) interval, but the log odds of a correct item response. The interval properties of logits are established by the joint relationship between person ability and item difficulty. Indeed, it would be possible for the logit scale to be interval according to conjoint measurement principles even if θ itself

[1] Though certainly not universal. For example, in his monograph "Rasch Models for Measurement," Andrich (1988) explicitly rejects the use of Steven's scale classifications.

[2] This might be a bit of an overstatement is the sense that many psychometricians may fall more squarely in a third school of thought that Michell refers to as the operational theory of measurement. According to operational theory, "the numbers don't remember where they came from" hence there is no rule governing the use of arithmentic operations as a function of scale type. In operationalism, measurement is defined by Stevens's dictum that measurement is "the assignment of numerals to objects or events according to rules." Such a definition (which ignores much of Stevens's representational perspective) is broad enough to encompass almost any formal measurement procedure. One might say that operationalists are defined more by what they do not endorse as a strict definition of requirements for measurement than what they do.

were only ordinal. However, it is possible to express a test respondent's value of θ in terms of its location on the logit scale. Because of this, when item responses meet the requirements of the Rasch Model, and when respondents and test items cover the full range of the ability distribution (this second condition is often overlooked), then one can support the claim that IRT produces a scale (logits) with interval properties. It would seem to follow from this that whether one is an adherent of classical or representational measurement theory, if a scale is desired for which magnitude should have consistent meaning, then the psychometric focus of test development should be to create tests that satisfy the constraints of the Rasch Model[3]. This does not seem to be the case in practice.

Our premise in this paper is that the conventional wisdom among psychometricians responsible for the development of large-scale tests for state testing programs is that score scales are ordinal, not interval. We suspect that the reason for this is the tenet, consistent with representational measurement theory, that the job of the psychometrician is to take the data she/he is given and make the most of it. One cannot weave gold from straw. Empirically, test items do not discriminate equally; they are not distributed along the full theta scale; they may not be unidimensional (especially when they are being linked vertically). More to the point, even if the assumptions of the Rasch Model were met, there is no satisfactory way to validate that conjoint additivity has been established. (Ballou takes this one step further by questioning whether interval scales established through conjoint additivity are even theoretically appealing relative to other

[3] Ballou suggests that interval properties for the 2PL and 3PL models could also be supported (in theory) if polynomial conjoint measurement could be established. This essentially boils down to restricting the range of the scale to a span that would support invariant person by item comparisons (for example, the range of the scale where ICCs do not cross).

possible alternatives.) Given all this, the most pragmatic approach is to proceed under the assumption that IRT scales are ordinal in nature.

There is nothing objectionable with this approach if the way that test score scales are being used to compute statistics is consistent with their fundamental properties. In this sense, the classification of students into performance levels as mandated under NCLB is entirely consistent with a belief that what tests measure is ordinal, not interval. But what might happen when psychometricians and state assessment directors proceed as if the scale was ordinal while policymakers and applied researchers use and interpret the scale as though it was interval? This seems particularly salient in the context of growth modeling. If the scale is ordinal, then any monotonic transformation of the scale is considered admissible. This appears to be why, in the context of vertical scaling applications, both Kolen & Brennan (2004) and Harris (2007) make the point that the concept of growth has no empirical definition, but can only be established externally[4]. Through the use of monotonic transformations, both the magnitude and variability of grade to grade scale score changes can be manipulated to take on any desired pattern (i.e., large gains and constant variance, small gains and increasing variance, etc.). Ballou (2008) suggests that this may make the estimation and interpretation of value-added quantities decidedly equivocal. In what follows we put this conjecture to an empirical test.

[4] This second assertion (that growth must be established externally through some theory of how student learn in each subject area) seems to contradict the philosophy of representational measurement theory that theory should follow from data and not the other way around. In spirit it is seems more closely aligned to a classical theory of measurement!

Methods

Data

The data we use in this study are longitudinal item responses from the Colorado Students Assessment Program (CSAP) reading test. The CSAP reading test has a vertical scale based on a common item nonequivalent groups linking design that was established by the state's test contractor, CTB/McGraw-Hill, in 2001. The longitudinal score scales used in the present study derive from data we obtained from the Colorado Department of Education for two cohorts of students. 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. 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 3PLM/GPCM IRT models, separate calibration and ML estimation (for details see Briggs & Weeks, in press). This scale, which we labeled "SEP3ML" serves as the base scale in our analyses. All subsequent scales are monotonic transformations of this base scale. All scales are expressed in the logit metric and are applied only to those students in cohort 1 for our subsequent growth modeling comparisons.

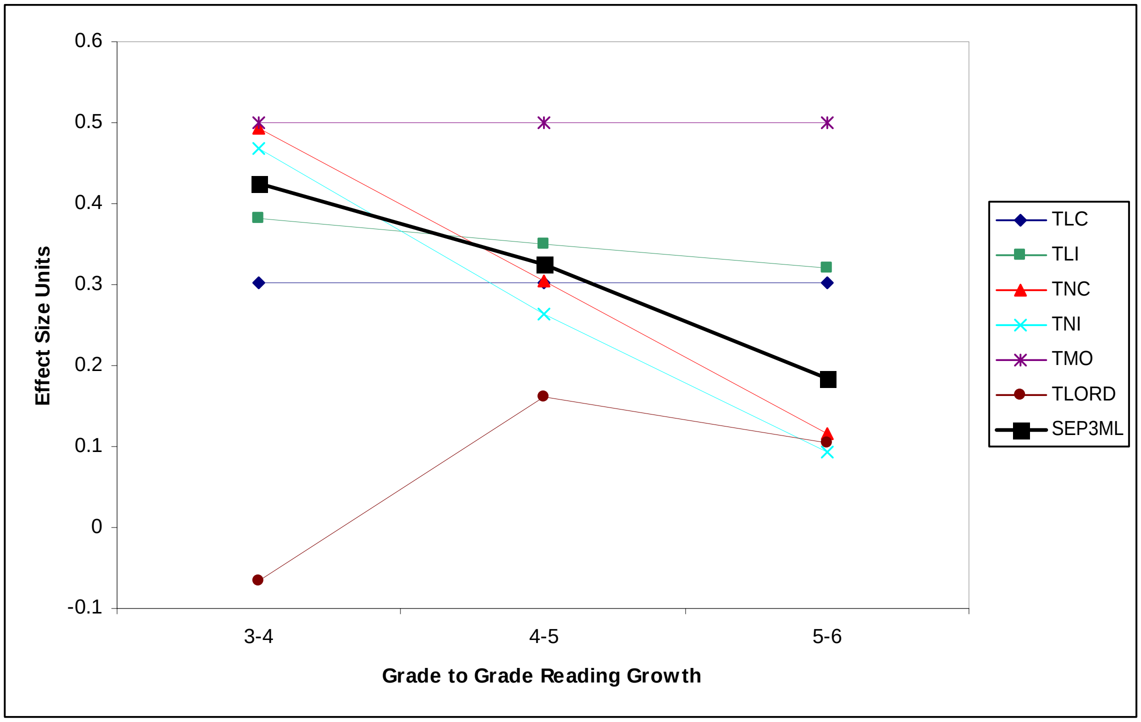
Monotonic Scale Transformations

We applied three types of monotonic transformations to our base scale (SEP3ML).

1. Minor adjustments to means and SDs by grade
    a. Linear growth and constant SD (TLC)
    b. Linear growth and increasing SD (TLI)
    c. Nonlinear growth and constant SD (TNC)
    d. Nonlinear growth and increasing SD (TNI)
2. More extreme adjustment to means and SDs by grade to decrease overlap in score distributions. (TMO)
3. Lord's transformation: exp(SEP3ML) = (TLORD)

The scales were transformed in approaches 1a-1d through the following process. First, the scores were multiplied by a grade-specific numeric factor so that the standard deviation of scores within each grade remained constant (at 1 logit) or increased (by .1 logits). Second, a grade-specific numeric factor was added to each score. This latter adjustment was derived by regressing longitudinal scores on time (linear case) or on time and time squared (nonlinear case). In approach 2, we set growth to be linear with a constant SD, but forced the amount of growth to be .5 logits (half of an SD) per year. Approaches 1 and 2 represent the kinds of policy-based scale transformations that have either occurred at the state level, or would not be considered surprising if they did occur. They would be "admissible" transformations if one only believes the underlying scale is ordinal rather than interval. Approach 3 represents our attempt to "break" the scale. Unlike the transformations in approaches 1 and 2, which were done so that grade to grade means and SDs would follow a pattern that "looks" right, the exponential transformation in approach 3 is based solely on Lord's (1980) observation that such a transformation could be applied to the theta metric while maintaining the same item response probabilities.

Figure 1. Comparison of Scale Transformations in Effect Size Units



Yen (1986) defined an effect size statistic for purposes of comparing year to year growth along a vertical scale as

$$\text{Effect Size} = \frac{\bar{\theta}_{upper} - \bar{\theta}_{lower}}{\sqrt{\frac{\sigma^2_{upper} + \sigma^2_{lower}}{2}}},$$

where $\bar{\theta}_{upper}$ and $\bar{\theta}_{lower}$ represent the mean scale scores for the higher and lower grades or years in the scale respectively, and σ^2_{upper} and σ^2_{lower} represent the respective variance for the scores in each grade or year. Figure 1 plots the effect sizes for the different scales we created. Note that in an absolute sense, these scales tell very different stories about student growth in reading, and each story is “admissible” if the underlying scale is ordinal. The question of interest here is whether such differences will have an impact on the normative interpretations of growth in student achievement when aggregated at the school-level.

Value-Added Estimates from the Layered Model

Our first model is a constrained version of the layered model (i.e., TVAAS/EVAAS) described by Sanders et al. (1997) and McCaffrey, et al. (2004). The model is constrained because it only uses a single longitudinal cohort and one test subject (reading). There is a further constraint in the sense that Lockwood et al. (2007) refer to the layered model as a “complete persistence” model because it assumes that teacher and/or school effects[5] remain constant over time. Note that because the model only considers longitudinal data for a single cohort of students, in the present context a “school effect” and a “grade effect” are the same thing. Applying it to each of the transformed scales we created for the time period from 2003 to 2006 yields the following system of equations

$$\begin{aligned}
Y_{i03} &= \mu_{03} + \boldsymbol{\theta}_{03} + \varepsilon_{i03} \\
Y_{i04} &= \mu_{04} + \boldsymbol{\theta}_{03} + \boldsymbol{\theta}_{04} + \varepsilon_{i04} \\
Y_{i05} &= \mu_{05} + \boldsymbol{\theta}_{03} + \boldsymbol{\theta}_{04} + \boldsymbol{\theta}_{05} + \varepsilon_{i05} \\
Y_{i06} &= \mu_{06} + \boldsymbol{\theta}_{03} + \boldsymbol{\theta}_{04} + \boldsymbol{\theta}_{05} + \boldsymbol{\theta}_{06} + \varepsilon_{i06}.
\end{aligned} \qquad (1)$$

[5] The term “residual” is actually more appropriate characterization of $\boldsymbol{\theta}_t$ than the term “effect,” but we use the latter to be consistent with the literature.

In the equations above, the $Y_{i's}$ represents the CSAP reading test score for student *i* in a given year and grade, and $\mu's$ denotes the state test score mean for a given year. (By coincidence that last digit in each year corresponds to the associated grade level of a student in the longitudinal cohort.) Each year/grade-specific $\boldsymbol{\theta}$ represents a vector of school effects. Finally, the $\varepsilon_i's$ represents the test score residual associated with student *i* in a given year. Both $\boldsymbol{\theta}_t$ and ε_{it} are assumed to be independent latent random variables (where the subscript *t* indexes year), where $\varepsilon_{it} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\theta}_t \sim N(0, \tau)$. These equations are linked together through the intercorrelations of the $\varepsilon_{i's}$. We note the following about the school-level parameters in these equations.

- First, the parameter vectors $\{\boldsymbol{\theta}_{04}, \boldsymbol{\theta}_{05}, \boldsymbol{\theta}_{06}\}$ represent the value-added by schools to the achievement of students in grades 4, 5 and 6 respectively. This is in contrast to the parameter vector $\boldsymbol{\theta}_{03}$, which captures pre-existing differences in school status as of grade 3. These parameters are specified in the model under the constraint that they are independent across years.
- Second, while the model above can be easily extended to allow for multivariate test outcomes (typical of applications of the EVAAS by Sanders), background covariates, and a term that links school effects to specific students in the event that students attend more than one school in a given year (c.f., Lockwood et al., 2007, p. 127-128), we have chosen this simpler specification in order to focus attention on the relationship between differences in our choice of the underlying scale and the resulting schools effect estimates.

- Third, we obtain estimates for our school-level parameters via Bayesian estimation procedures using an application developed by J. R. Lockwood and described in Lockwood et al. (2007). For each school in a grades 4 through 6, we are able to estimate a posterior distribution of the school’s value-added effect on student reading performance. We subsequently use the mean of this posterior distribution as a point estimate for this effect. Value-added effect have a normative interpretation in the layered model, and can be interpreted as the deviation from the average Colorado public school. Finally, because many students in Colorado transition from elementary school to middle school after grade 5, the total number of schools for which effects are estimated decreases from 950 to 640 as of 2006.

The layered model is a multivariate mixed effects model, and in some sense it straddles the fence in our distinction between models of absolute and relative growth. Because it does not condition on prior achievement within any of its grade-specific layers, it can be conceptualized as an absolute growth model. For this reason, as Ballou et. al (2004) note, the layered model appears to require longitudinal test scores that have been vertically scaled. On the other hand, the layered model is solely used to produce estimates of school or teacher-level value-added, quantities that have a purely relative, norm-referenced interpretation.

Student-Growth Percentiles (SGPs) from Quantile Regression

Our second model is a relative growth model currently used by Colorado and Massachusetts to report student growth and approved for Colorado by the U.S.

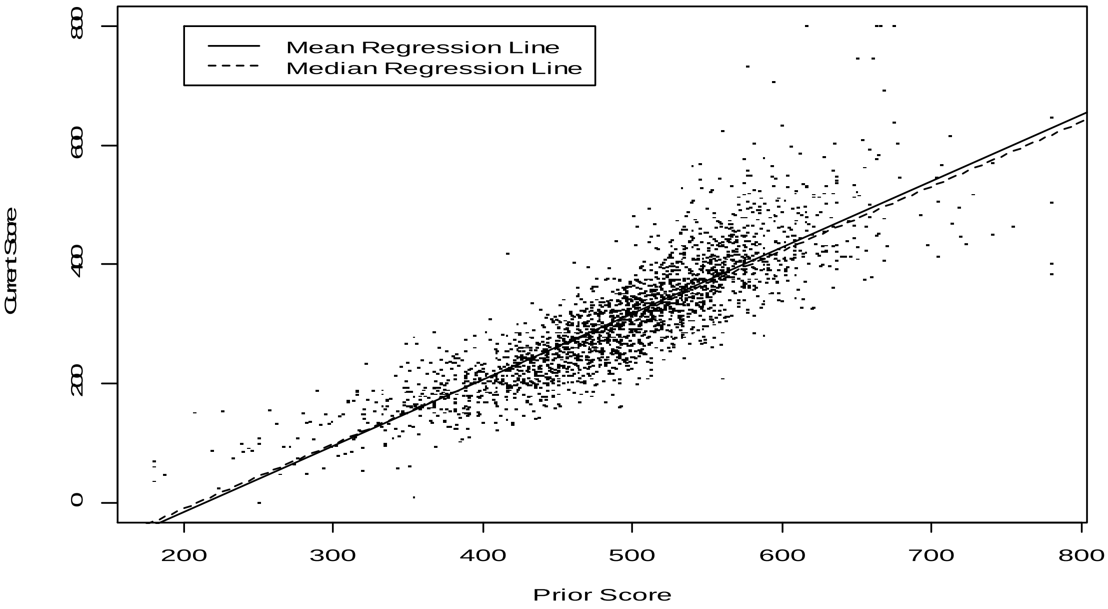
Department of Education as part of the Growth Model Pilot Program. To quantify student growth normatively, a student's current score is located within the conditional distribution of current scores based upon all prior scores to give a *student growth percentile:* The percentile of a student's current score within their corresponding conditional distribution translates to a probability statement of a student obtaining that score taking account of prior achievement. That is:

Student Growth Percentile = Pr(Current Achievement | Past Achievement) · 100.

Calculation of a student's growth percentile is based upon the estimation of the conditional density associated with a student's score at time t using the student's prior scores at times 1, 2, . . . , $t - 1$ as the conditioning variables. Given the conditional density for the student's score at time t, the student's growth percentile is defined as the percentile of the score within the time t conditional density. By examining a student's current achievement with regard to the conditional density, the student's growth percentile normatively situates the student's outcome at time t taking account of past student performance. The percentile result reflects the likelihood of such an outcome given the student's prior achievement. In the sense that the student growth percentile translates to the probability of such an outcome occurring (i.e., rarity), it is possible to compare the progress of individuals not beginning at the same starting point. However, occurrences being equally rare do not necessarily imply that they are equally "good". Qualifying student growth percentiles as "good", "(in)adequate", or as satisfying "a year's growth" is a standard setting procedure requiring external criteria (e.g., growth relative to state performance standards) combined with the wisdom and judgments of stakeholders.

Estimation of the conditional density is performed using quantile regression (Koenker, 2005). Whereas linear regression methods model the conditional mean of a response variable *Y*, quantile regression is more generally concerned with the estimation of the family of conditional quantiles of *Y*. The simplest example involves the estimation of the median regression line (quantile =0.5). This line models the conditional median of the response variable instead of the conditional mean. Figure 2 shows a scatterplot with both the mean and median regression lines:

Figure 2. Illustration of Two Different Regression Lines: Conditional Mean vs. Median



In addition to the median regression line, regression lines for each quartile, decile, or any quantile can be examined. As such, quantile regression provides a more complete picture of both the conditional distribution associated with the response variable.

The techniques are ideally suited for estimation of the family of conditional quantile functions. Using quantile regression, the conditional density associated with each student's prior scores is derived and used to situate the student's most recent score. Position of the student's most recent score within this density can then be used to qualify deficient/sufficient/excellent growth. Though many state assessments possess a vertical scale, such a scale is not necessary to calculate student growth percentiles.

To accommodate non-linearity, heteroskedasticity, and skewness of the distribution, non-parametric B-spline smoothing is employed. B-splines are attractive both theoretically and computationally in that they are appropriate given the known patterns of variability along test scales, seldom lead to estimation problems (Harrell, 2001, p. 20), and are simple to implement in available software. Calculation of student growth percentiles is performed using R (R Development Core Team, 2009), a language and environment for statistical computing, with the SGP package (Betebenner, 2009). Other possible software (untested with regard to student growth percentiles) with quantile regression capability include SAS and Stata.

For the present data, student growth percentiles (SGPs) are computed for the students in our longitudinal cohort separately for three grades: grade 4 (conditioning on grade 3), grade 5 (conditioning on grades 3 and 4), and grade 6 (conditioning on grades 3, 4 and 5). These SGPs are then aggregated to the school-level by taking the median. We do not refer to school-level SGPs as value-added estimates for two reasons. First, no residual has been computed (though this could be done easily enough by subtracting the 50th percentile), and second, we wish to avoid the causal inference that high or low SGPs can be explained by high or low school quality (for details, see Betebenner, 2008).

Results

We organize our presentation of results as follows. First, we compare the correlations of school-level estimates by grade across our seven different test scales *within* our two growth modeling approaches. Next, we correlate and plot the school-level estimates *across* growth models by grade holding the underlying base scale constant. In both our within and across model comparisons we also present correlations of value-added estimates with two key school by grade specific measures of status: the percentage of students eligible for free or reduced price school lunches (FRL), and the mean of prior year reading achievement on the CSAP A big part of the theoretical appeal of value-added estimates is that they should be much less correlated (some would even claim uncorrelated) with these sorts of status measures. Of interest is whether and to what extent the correlations between value-added and status vary by scale and growth model. Finally, we include correlations with the mean of current year reading achievement on the CSAP. There is no reason to expect value-added estimates to be uncorrelated with this measure (in fact, one would expect a positive correlation since schools with higher student growth/value-added over the last year should, on average, demonstrate high achievement at the end of the growth/value-added cycle), but it serves as a useful descriptive statistic.

Comparisons Across Scales Within Growth Model

The theory behind the use of quantile regression provides invariance to monotonic transformation of scale with regard to the dependent variable (Koenker, 2005, p. 39). With regard to transformations across both independent and dependent variables, the expectation is that it will be invariant to such transformations of the underlying score scale. Our empirical results support this. In almost all cases the median school-level SGPs in each grade were perfectly correlated across the seven different scales we created. The one exception were correlations with the scale created by applying the exp transformation (TLORD), but even in this case median SGPs correlated .98 with SGPs based on the other scales. In addition, the correlations between SGPs and school-level status measures also remain constant across scales. We conclude that it is safe to say that the interpretations of SGPs are not scale-dependent.

The correlation between median SGP and percent free/reduced lunch is moderately negative across the three year/grade analyses ranging from -0.42, -0.25, to -0.39 in grade/year 4/2004, 5/2005, and 6/2006, respectively. Note that when prior (or current) achievement is correlated with free/reduced lunch percentage (without conditioning upon prior achievement), the correlations are very strong in absolute magnitude, more than -0.8 in almost all instances, confirming the long held understanding of the relationship between achievement and poverty.

Correlations between median SGP and prior achievement range from 0.34, 0.18, to 0.31 in grade/year 4/2004, 5/2005, and 6/2006, respectively. These result suggest that schools that higher achieving students tend to, on average, show higher normative rates of growth than schools serving lower achieving students. Making the “inferential leap” that student growth is solely caused by the school and sources of influence therein, the results translate to saying that schools serving higher achieving students tend to, on average, be more effective than schools serving lower achieving students. The correlations between median SGP and current achievement are (tautologically) higher reflecting the fact that students growing faster show higher rates of achievement that is reflected in higher average rates of achievement at the school level.

Table 1. Correlation Table of 2004, Grade 4, LM Value-Added Estimates by Scale, Free/Reduced Lunch Percentage, Prior and Current Mean Theta

	S3ML	**TLC**	**TLI**	**TNC**	**TNI**	**TMO**	**TLORD**	**FRL**	**PRIOR THETA**	**THETA 04**
SEP3ML	1.00	0.96	0.87	0.96	0.87	0.95	0.30	-0.15	-0.11	0.23
TLC	0.96	1.00	0.97	1.00	0.97	1.00	0.52	-0.38	0.17	0.49
TLI	0.87	0.97	1.00	0.97	1.00	0.97	0.66	-0.52	0.38	0.67
TNC	0.96	1.00	0.97	1.00	0.97	1.00	0.52	-0.38	0.17	0.49
TNI	0.87	0.97	1.00	0.97	1.00	0.97	0.66	-0.52	0.38	0.67
TMO	0.95	1.00	0.97	1.00	0.97	1.00	0.52	-0.38	0.17	0.49
TLORD	0.30	0.52	0.66	0.52	0.66	0.52	1.00	-0.68	0.78	0.87
FRL	-0.15	-0.38	-0.52	-0.38	-0.52	-0.38	-0.68	1.00	-0.77	-0.81
PRIOR THETA	-0.11	0.17	0.38	0.17	0.38	0.17	0.78	-0.77	1.00	0.93
THETA04	0.23	0.49	0.67	0.49	0.67	0.49	0.87	-0.81	0.93	1.00

Table 2. Correlation Table of 2005, Grade 5, LM Value-Added Estimates by Scale, Free/Reduced Lunch Percentage, Prior and Current Mean Theta

	S3ML	**TLC**	**TLI**	**TNC**	**TNI**	**TMO**	**TLORD**	**FRL**	**PRIOR THETA**	**THETA 05**
SEP3ML	1.00	1.00	0.95	1.00	0.95	1.00	0.45	-0.06	-0.09	0.20
TLC	0.96	0.95	0.85	0.95	0.85	0.95	0.36	0.12	-0.27	0.00
TLI	1.00	1.00	0.96	1.00	0.96	1.00	0.48	-0.10	-0.04	0.24
TNC	0.95	0.96	1.00	0.96	1.00	0.96	0.59	-0.33	0.22	0.49
TNI	1.00	1.00	0.96	1.00	0.96	1.00	0.48	-0.10	-0.04	0.24
TMO	0.95	0.96	1.00	0.96	1.00	0.96	0.60	-0.33	0.22	0.49
TLORD	1.00	1.00	0.96	1.00	0.96	1.00	0.47	-0.10	-0.04	0.24
FRL	0.45	0.48	0.59	0.48	0.60	0.47	1.00	-0.47	0.48	0.60
PRIOR THETA	-0.06	-0.10	-0.33	-0.10	-0.33	-0.10	-0.47	1.00	-0.86	-0.86
THETA05	-0.09	-0.04	0.22	-0.04	0.22	-0.04	0.48	-0.86	1.00	0.95

Table 3. Correlation Table of 2006, Grade 6, LM Value-Added Estimates by Scale, Free/Reduced Lunch Percentage, Prior and Current Mean Theta

	S3ML	**TLC**	**TLI**	**TNC**	**TNI**	**TMO**	**TLORD**	**FRL**	**PRIOR THETA**	**THETA 06**
SEP3ML	1.00	0.98	1.00	0.98	1.00	0.98	0.51	-0.51	0.39	0.62
TLC	0.98	1.00	0.98	1.00	0.98	1.00	0.43	-0.36	0.22	0.47
TLI	1.00	0.98	1.00	0.98	1.00	0.98	0.50	-0.50	0.39	0.62
TNC	0.98	1.00	0.98	1.00	0.98	1.00	0.43	-0.36	0.22	0.47
TNI	1.00	0.98	1.00	0.98	1.00	0.98	0.50	-0.50	0.39	0.62
TMO	0.98	1.00	0.98	1.00	0.98	1.00	0.43	-0.36	0.22	0.47
TLORD	0.51	0.43	0.50	0.43	0.50	0.43	1.00	-0.48	0.52	0.58
FRL	-0.51	-0.36	-0.50	-0.36	-0.50	-0.36	-0.48	1.00	-0.84	-0.87
PRIOR THETA	0.39	0.22	0.39	0.22	0.39	0.22	0.52	-0.84	1.00	0.96
THETA06	0.62	0.47	0.62	0.47	0.62	0.47	0.58	-0.87	0.96	1.00

Much of the comparisons across scales within the layered model also indicate a lack of sensitivity to the monotonic scale transformations we consider. The correlations across scales (shown in Tables 1-3) tend to be very strong with the notable exception of the TLORD scale. Value-added estimates based on the TLORD scale tended to be weakly to moderately correlated with estimates based on the other six scales. Again, there is no good reason why one would create a scale by applying this sort of transformation, but our intent was to see whether an extreme transformation could be found that would lead to distortions in the value-added estimates from the layered model, and in this sense we were successful.

It is not clear from our analyses whether the lack of perfect correlation between the scales is due to the transformation altering assumptions necessary for linearity to hold over time, altering the scale such that whatever interval nature is assumed changes dramatically from scale to scale, or for some other reason. In applications of such models it is worthwhile to know how violations of assumptions can lead to different results. We hope to tease out these factors in greater detail in later research.

One interesting pattern (consistent with previous finding by Briggs & Weeks) is that scale transformations that involved increases to the grade to grade variability of the base scale (TLI and TNI) led to small decreases in subsequent correlations of the associated value-added estimates across other scales in which variability was held constant (TLC, TNC, TMO) or was estimated empirically (SEP3ML). An unexpected finding with respect to the layered model is that scale transformations appear to have an impact on the correlation with measures of school-level status. For example, in grade 4 the correlation of value-added estimates deriving from our base scale with FRL% was $-.15$. This correlation shifts to $-.38$ under the TLC, TNC and TMO scales and $-.52$ under the TLI and TNI scales. A similar pattern is found for the TLI and TNI results in grade 5 (though not for the TLC, TNC and TMO scales). As of grade 6, all scales (including the base scale) exhibit a weak to moderate negative correlation with FRL%, but we note that the pattern of differential correlations as a function of transformation to scale variability are still evident. When grade to grade scale variability is increased, correlations of grade-specific value-added estimates with measures of school-level status also increase in absolute magnitude.

Table 4. 2004 Correlations of School-level Estimates by Model & Status Measures (Schools >= 50 Students)

	SGP	**LM**	**FRL_2004**	**PRIOR THETA**	**CURRENT THETA**
SGP	1.00	0.72	-0.42	0.34	0.59
LM	0.72	1.00	-0.15	-0.11	0.23
FRL_2004	-0.42	-0.15	1.00	-0.77	-0.81
PRIOR THETA	0.34	-0.11	-0.77	1.00	0.93
CURRENT THETA	0.59	0.23	-0.81	0.93	1.00

Note: School-level estimates from use of base scale (SEP3ML) for CSAP Reading Test, N = 557

Table 5. 2005 Correlations of School-level Estimates by Model & Status Measures (Schools >= 50 Students)

	SGP	**LM**	**FRL_2005**	**PRIOR THETA**	**CURRENT THETA**
SGP	1.00	0.84	-0.25	0.18	0.42
LM	0.84	1.00	-0.06	-0.09	0.20
FRL_2005	-0.25	-0.06	1.00	-0.86	-0.86
PRIOR THETA	0.18	-0.09	-0.86	1.00	0.95
CURRENT THETA	0.42	0.20	-0.86	0.95	1.00

Note: School-level estimates from use of base scale (SEP3ML) for CSAP Reading Test, N = 569

Table 6. 2006 Correlations of School-level Estimates by Model and Status Measures (Schools >= 50 Students)

	SGP	**LM**	**FRL_2006**	**PRIOR THETA**	**CURRENT THETA**
SGP	1.00	0.91	-0.39	0.31	0.52
LM	0.91	1.00	-0.51	0.39	0.62
FRL_2006	-0.39	-0.51	1.00	-0.84	-0.87
PRIOR THETA	0.31	0.39	-0.84	1.00	0.96
CURRENT THETA	0.52	0.62	-0.87	0.96	1.00

Note: School-level estimates from use of base scale (SEP3ML) for CSAP Reading Test, N = 380

Table 7. 2004 Correlations of Grade 4 School-level Estimates by Model and Status Measures (All Schools)

	SGP	**LM**	**FRL_2004**	**PRIOR THETA**	**CURRENT THETA**
SGP	1.00	0.69	-0.31	0.25	0.59
LM	0.69	1.00	-0.13	-0.09	0.30
FRL_2004	-0.31	-0.13	1.00	-0.68	-0.69
PRIOR THETA	0.25	-0.09	-0.68	1.00	0.89
THETA04	0.59	0.30	-0.69	0.89	1.00

Note: School-level estimates from use of base scale (SEP3ML) for CSAP Reading Test, N = 940

Table 8. 2005 Correlations of Grade 5 School-level Estimates by Model and Status Measures (All Schools)

	SGP	**LM**	**FRL_2005**	**PRIOR THETA**	**CURRENT THETA**
SGP	1.00	0.79	-0.18	0.14	0.43
LM	0.79	1.00	-0.05	-0.10	0.21
FRL_2005	-0.18	-0.05	1.00	-0.75	-0.73
PRIOR THETA	0.14	-0.10	-0.75	1.00	0.93
CURRENT THETA	0.43	0.21	-0.73	0.93	1.00

Note: School-level estimates from use of base scale (SEP3ML) for CSAP Reading Test, N = 948

Table 9. 2006 Correlations of Grade 6 School-level Estimates by Model and Status Measures (All Schools)

	SGP	**LM**	**FRL_2006**	**PRIOR THETA**	**CURRENT THETA**
SGP	1.00	0.82	-0.24	0.25	0.51
LM	0.82	1.00	-0.35	0.29	0.53
FRL_2006	-0.24	-0.35	1.00	-0.68	-0.68
PRIOR THETA	0.25	0.29	-0.68	1.00	0.92
CURRENT THETA	0.51	0.53	-0.68	0.92	1.00

Note: School-level estimates from use of base scale (SEP3ML) for CSAP Reading Test, N = 637

Comparisons Across Growth Models Within Base Scale

Tables 4-6 and 7-9 present the correlations of school-level estimates deriving from the SGP and the layered model (LM) by grade. In tables 4-6 these correlations are based only upon schools with at least 50 students tested in a given grade. Tables 7-9 include all schools regardless of the number of tested students. In tables 4-6 the correlations of school-level estimates between the two models are .72, .84 and .91 for grades 4, 5 and 6 respectively. When all schools are considered (tables 7-9), the respective correlations decrease slightly to .69, .79 and .82. Figures 3a-3c (included as a separate document) give a graphical depiction of the relationship between SGP and LM estimates for each grade. The size of each "bubble" in the plot represents the number of students in a given school in that grade. The shade of each bubble indicates the school-level quintile of prior reading achievement.

The plots are divided into four quadrants. The quadrants of greatest interest are II and IV. Schools landing in these quadrants exhibit growth that is better than expected under one model but worse than expected under the other. These are examples of potential classification errors. It is primarily for the grade 4 estimates (where the correlations between LM estimates and SGPs are lowest) that we see a cluster of unusual estimates in quadrant II. These are schools that appear to be relatively high-growth using SGPs but low-growth using the LM. Interestingly they tend to be schools in the highest grade 3 achievement quintile.

We note that in general across all three plots, there are many examples of small schools with rather high or low SGPs (relative to what would be expected for the average school—the median SGP) that have LM estimates of value-added that are much less extreme (again relative to what would be expected for the average school—the state average of residuals, 0). This is a reflection of the fact that the LM produces shrunken value-added estimates while median SGP are not shrunk. This distinction gets to the heart of what differs between a growth model and a value-added model. The median SGP describes the growth of a "typical" student without ascribing responsibility (i.e., value-added) for that growth. The students, in effect, are fixed. The shrunken VAM estimate reflects the uncertainty associated with the attribution of effectiveness to the unit based upon so few observations. As Koenker notes (Koenker, 2005, p. 278), shrinkage ultimately has the effect of adding pseudo-observations to the fixed effects estimator that, in effect, pulls the result toward the grand mean indicating the lack of certainty associated with effectiveness being any different than the average effectiveness found in the district.

We now turn to the correlations of school estimates with our three school-level measures of status. In grades 4 and 5 there is a fairly clear pattern in model comparisons. Value-added estimates based on both quantile regression and the LM tend to both be weakly correlated with FRL% and prior achievement, and the correlations are always larger for SGPs. For grade 6 this pattern changes: under the LM the value-added correlations with FRL% and prior achievement become considerably stronger—in fact, the correlations become larger in magnitude for the LM than for the quantile regression model. In contrast, note that the correlations for the QR model stay for the most part consistent with those found in grades 4 and 5. This is true whether we consider value-added estimates based on only those schools with more than 50 tested students, or on all available schools.

Discussion

Is growth in student achievement scale dependent? If growth is understand in terms of magnitude then the magnitude of student growth is scale dependent. When growth is understood normatively, then the answer is equivocal. We have shown two modeling approaches where growth interpretations appear (for the most part) insensitive to reasonable monotonic transformations to the underlying score scale. In the case of the SGPs produced by QR, this is a theoretical feature of the model (only an ordinal score scale for the outcome variable is required), and we have demonstrated this feature empirically. For the LM, on the other hand, a vertical scale with interval properties is generally considered to be a requirement. However, in this empirical context, we find that the ordering of school-level value-added estimates was largely insensitive all our scale transformations (with the exception of the TLORD scale). The only place where we see sensitivity to the choice of scale for the LM is in the correlation between value-added estimates and measures of school-level status. In particular, there appears to be an interaction between the underlying scale transformation and the grade level for which value-added school effects have been estimated with the present data. We are still in the process of investigating this interaction.

Both the QR approach and LM share something in common: both approaches produce quantities that provide for normative, rather than criterion-referenced interpretations of growth. Both models allow us to assess whether conditional achievement is “good enough” from a norm-referenced perspective. However, there is an important philosophical difference between the two modeling approaches in that Betebenner (2008) has focused upon the use of SGPs as a descriptive tool to characterize growth at the student-level, while the LM is typically the engine behind the teacher or school effects that get produced for inferential purposes in the EVAAS. Nonetheless, in this application we find that the correlations between school-level SGPs from QR and value-added estimates from the LM tend to be strongly correlated. This correlation of grade-specific school effects grows stronger as more prior test scores are used as conditioning variables in the QR used to estimate SGPs. Note that in contrast to QR, the LM "adjusts" estimated effects in earlier grades on the basis of both prior *and* subsequent test score performance of students. This is why, in principle, one would expect, LM value-added estimates to have a weaker correlation with school-level measures of status than that found with QR SGP estimates. This is generally the case here with the exception of the anomalous findings in grade 6.

Given our results, should vertical score scales with interval properties be considered a desirable and/or necessary feature of an accountability system hoping to answer questions about growth? Our take is that if we wish to answer intuitively meaningful questions about how *much* a student has learned, then a strong argument can be made that vertical scales with interval properties are both desirable and necessary. For models of absolute growth in which criterion-referenced interpretations are of interest, vertical scales are central to the endeavor. In this context it is useful to remember the other term often used synonymously with vertical scales: *developmental* scales. However, given the vertical scales currently in use, their existence is far from sufficient

Unfortunately at present we see limited evidence that states and their test contractors are working in tandem to create test score metrics that could plausibly measure and communicate information about magnitudes of student learning. One obstacle to this might be a failure of the psychometric community to have sustained and serious dialogues about the premises under which score scales are being developed (Michell, 2000). If the underlying scale is only ordinal, then it may not be reasonable to compare means and SDs by subgroups, or to use longitudinal data with models of absolute growth. If an interval scale is essential to the planned use of test scores, than we see no workable way around the need for a Rasch family IRT model. The present state of affairs among vertical scales in large-scale assessment has brought us to a point where statements about magnitudes of criterion-referenced growth are largely meaningless. But just because something has not been done well does not mean that it can't be done nor should not be done at all.

References

Andrich, D. (1988) Rasch models for measurement. Beverly Hills: Sage Publications.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37.

Ballou, D. (2008). Test Scaling and Value-Added Measurement. Presented at *National Conference on Value-Added Modeling,* Madison, Wisconsin.

Betebenner, D. W. (2009). SGP: Student growth percentile and percentile growth projection/trajectory functions [Computer software manual]. (R package version 0.0-4)

Briggs, D. & Weeks, J. (in press). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy*.

Cronbach, L. J., & Furby, L. (1970). How should we measure" change"—or should we. *Psychological Bulletin*, *74*(1), 68-80.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 445-492.

Harrell, F. E. (2001). Regression modeling strategies. New York: Springer.

Harris, D. (2007). Practical issues in vertical scaling In *Linking and aligning scores and scales*. N. J. Dorans, M. Pommerich, & P.Holland (editors). Springer.

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

Koenker, R. (2005). Quantile regression. Cambridge: Cambridge University Press.

Lissitz, R. W. (2005). *Value added models in education: Theory and applications*. Jam Press.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, *32*(2), 125.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum Associates, Hillsdale, NJ.

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

McCaffrey, D. F. , Lockwood, J. R, Koretz, D., Louis, T. A, and Hamilton, L. (2004) Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, Vol. 29:1, 67-101.

McCaffrey, D., Han, B., & Lockwood, J. (2008). From data to bonuses: A case study of the issues related to awarding teachers pay on the basis of their students' progress. In *National Center for Performance Incentives conference Performance Incentives: Their Growing Impact on American K-12 Education, Vanderbilt University, Nashville, TN.*

Michell, J. (1986). Measurement scales and statistics: a clash of paradigms. *Psychological Bulletin,* 100, 398-407.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Lawrence Erlbaum.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory and Psychology*, *10*(5), 639-668.

R Development Core Team. (2009). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org (3-900051-07-0)

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage Publications Inc.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*(3), 726-748.

Ryan, K., & Shepard, L. A. (2008). *The future of test-based educational accountability*. Routledge.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Accountability System: A quantitative, outcomes-based approach to educational assessment. *Grading teachers, grading schools: Is student achievement a valid evaluation measure*, 137–162.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press, USA.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680.

Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin*, *96*(2), 394-401.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In *Test scoring*, D. Thissen & H. Wainer (eds), 73–140.

Willett, J. B. (1988). Chapter 9: Questions and Answers in the Measurement of Change. *Review of research in education*, *15*(1), 345.

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

Zwick, R. (1992). Statistical and Psychometric Issues in the Measurement of Educational Achievement Trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Statistics*, *17*(2), 205-218.