The Sensitivity of Value-Added Modeling to the Creation of a Vertical Score Scale

Derek C. Briggs

Jonathan P. Weeks

University of Colorado, Boulder

February, 2009

Pre-print for Briggs, D. C. & Weeks, J. P. (2009) The sensitivity of value-added modeling to the creation of a vertical scale. *Education Finance & Policy*, 4(4), 384-414.

The research described in this paper was supported by a grant from the Carnegie Corporation.

Abstract

The purpose of this study was to evaluate the sensitivity of growth and value-added modeling to the way an underlying vertical score scale has been created. Longitudinal item-level data was analyzed with both student and school-level identifiers for the entire state of Colorado between 2003 and 2006. Eight different vertical scales were established on the basis of choices made for three key variables: Item Response Theory modeling approach, calibration approach and student proficiency estimation approach. Each scale represented a methodological approach that was psychometrically defensible. Longitudinal values from each scale were used as the outcome in a commonly used value-added model (the "layered model" popularized by William Sanders) as a means of estimating school effects. Our findings suggest that while the ordering of estimating school effects is insensitive to the underlying vertical scale, the precision of such value-added estimates can be quite sensitive to the combinations of choices made in the creation of the scale.

Introduction

The idea of "value-added analysis" originated in the economics literature of the 1960s (Miller & Modigliani, 1961) but has more recently been used in education to characterize the added impact of teachers or schools on student gains relative to the gains students would be predicted to make with the average teacher or school respectively (McCaffrey, Lockwood, Koretz & Hamilton, 2003; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). Such uses of value-added models (VAMs) for the purposes of educational accountability (what Harris, 2008 has described as "VAM for Accountability") is the primary focus of this paper. As with any statistical model, VAMs can be written as an equation in which some measure of achievement (i.e., the left hand side of the equation) is expressed as a function of explanatory variables (i.e., the right hand side of the equation). Until recently, the bulk of the research literature on VAMs has been devoted to a careful consideration for how the right hand side of the equation should be specified: Should teacher effect parameters be specified such that they persist over time or should they be allowed to decay (McCaffrey et al., 2004)? Should student, teacher or school covariates be included (Ballou, Sanders & Wright, 2004)? Should value-added effects be modeled as fixed or random (Harris, 2008). Can value-added estimates be given a causal interpretation (Rubin, Stuart & Zannato, 2004; Raudenbush, 2004)? The focus of the present study is on the measurement specification of student achievement found on the left hand side of VAM equations.

As part of a comprehensive evaluation of VAMS, McCaffrey at al (2002) drew attention to number of psychometric issues fundamental to the construction of tests and scaling of test scores. In particular, they noted that there was reason to suspect that VAM estimates might be

sensitive to the ways that longitudinal test scores were linked across years. The purpose of a vertical scaling process is to ensure that the test scores for a given subject (i.e., reading, math, etc.) in earlier grades can be meaningfully compared to the test scores in later grades. The raw scores on such tests (i.e., proportion of items answered correctly) are clearly not comparable because the tests will necessarily differ with respect to their difficulty. For example, in an absolute sense a reading test in grade 3 will be easier than a reading test in grade 4, which will be easier than a test in grade 5, and so on. To place student performance onto a common scale, scores from two or more tests must be linked statistically to create what is known as a vertical scale. According to McCaffrey et al

> Changes to the scaling of tests, the weight given to alternative topics, or the methods for vertical linking could change our conclusions about the relative achievement or growth in achievement among classes of students…we expect 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. There is currently no empirical evidence about the sensitivity of gain score or teacher effects to such alternatives. (p. 89)

The purpose of this paper is to empirically evaluate the extent to which longitudinal interpretations of student score changes are sensitive to the decisions made in creating a vertically linked scale. We accomplish this by analyzing four years of longitudinal item-level reading data with both student and school-level identifiers for the entire state of Colorado. We use this data to address two principal research questions:

1. What is the sensitivity of a longitudinal score scale to the way test scores have been vertically linked?
2. What impact do different vertical scaling approaches have on subsequent estimates of

value-added school effects?

The basic strategy taken here is to create different vertical scales based on three key variables: the item response theory (IRT) model used to estimate item parameters, (2) the linking method used to place the parameters from different grades onto a common scale, and (3) the method used to estimate student-level scale scores. Combinations among these three variables lead to eight different vertical scales, each of which represents a methodological approach that is psychometrically defensible (i.e., there are no "straw man" scales). After creating the various scales, we first examine the patterns and differences in score means and standard deviations from year to year. Next, we treat the scores from each scale as outcome variables in a linear mixed effects model known as the "layered model[1]" (McCaffrey et al.., 2004; Sanders, Saxton & Horn, 1997). Of principal interest in this analysis are comparisons among estimates of school-level effects across scales.

Using Item Response Theory to Establish a Vertical Scale

Choosing an IRT Model

In IRT, an examinee's score on a test item is modeled probabilistically as a function of the examinee's latent proficiency and an item's characteristics. Let the variable X_{pi} represent the response of examinee p to item i. Given a test consisting of multiple-choice items, $X_{pi} = 1$ for a correct item response, and $X_{pi} = 0$ for an incorrect response. The item characteristic curve (ICC) for what is known as the three-parameter logistic model (3PLM: Birnbaum, 1968) can be written in the following form

$$P(X_{pi}=1\,|\,\theta_p,\alpha_i,\beta_i,\gamma_i)=\gamma_i+(1-\gamma_i)\frac{\exp(D\alpha_i(\theta_p-\beta_i))}{1+\exp(D\alpha_i(\theta_p-\beta_i))}. \quad (1)$$

In the formulation above, θ_p represents latent proficiency (measured in logits, which are the log of the odds of a correct item response), α_i represents item discrimination (slope of the ICC at the location of item difficulty), β_i represents item difficulty (value of θ_p at the ICC inflection point), γ_i represents a lower asymptote (pseudo-guessing parameter), D represents a scaling constant that equals 1.7, and "exp" represents the exponential function. When constraints are placed on the guessing and discrimination parameters in the 3PLM, two other commonly used models for dichotomously scored items can be derived. For the two-parameter logistic model (2PLM: Birnbaum, 1968) the guessing parameter γ_i is constrained to zero, and for the one-parameter model (1PLM) the 2PLM is constrained so that the discrimination parameters α_i are the same for all items. Though it has different historical and philosophical origins, mathematically the Rasch model (Rasch, 1960) can be viewed as a special case of the 1PLM where all of the item discriminations are constrained to equal one.

Insert Figure 1 about here

In an analogous fashion, constructed response items, generally referred to as polytomous items, can be modeled with what is known as the generalized partial credit model (GPCM: Muraki, 1992). The GPCM is akin to the 2PLM in a polytomous context (there is generally no guessing associated with constructed response items), and in the same way that the 2PLM can be constrained to produce the 1PLM, the discrimination parameters can be constrained to be equal for all polytomous items. This is known as the partial credit model (PCM: Masters, 1982). In

practice, when tests for large-scale assessment include a mixture of dichotomous and polytomous items they are commonly modeled using either a combination of the 3PLM and GPCM or the 1PLM and PCM.

One fundamental assumption all IRT models make is that of local independence, which posits that conditional on an examinee's latent proficiency, θ_p, item responses within a given test should be statistically independent observations. A related assumption is that the construct of measurement is unidimensional. Given that these assumptions are met and that the correct functional form has been specified, all IRT models share a very useful property: parameter invariance. According to this property, which is the linchpin for using IRT to establish a vertical score scale, item parameters are independent of the specific characteristics of the sample of test-takers used to estimate them. This implies that if 4th grade students and 5th grade students answer the same items on a reading test, the difficulty of the items should be the same regardless of which group was used to estimate them—even though in an absolute sense the items should generally be easier for the 5th grade students because they will likely have higher reading ability than 4th grade students.

Using IRT, two test score scales can be linked together provided that (a) the tests measure the same construct, and (b) the tests share a set of common items. This strategy constitutes what is formally known as a "common-item nonequivalent groups" (Kolen & Brennan, 2004), or a "non-equivalent groups with anchor test" (von Davier, Holland, & Thayer, 2004) linking design. The item response data we use in the present study are based upon this design[2].

Choosing a Linking Approach

In general, there are two IRT approaches used to create a vertical scale across two or more different grades: separate or concurrent linking[3]. Under a separate approach, item parameters and examinee ability (i.e., the achievement level that corresponds to an individual's scale score) are first estimated in separate runs for each grade-level test using the same IRT model and ability estimation approach (described below). Ultimately, the aim is to place the ability estimates from the various tests onto a common scale, and under the separate linking approach this is accomplished using a set of linear transformations. For example, imagine that two reading tests are administered at the end of a school year: one test is given to grade 4 students, the other to grade 5 students. The transformation $\theta^*_{p5} = A\theta_{p5} + B$ can be used to place the scores for 5th grade students on the 4th grade scale, where A and B are linking constants that adjust the standard deviation and mean ability of the 5th grade scores respectively, and θ^*_{p5} is the transformed score. If the same examinees take both tests, the linking constants can be readily computed, but such a design is not generally feasible. Within a common-item nonequivalent groups linking design, these constants can be found through a comparison of the differences between the estimated parameters for a set of common items on the two tests. Continuing with this example, imagine that each test has 50 multiple-choice items with 15 items that are common to both grades. In general, the items on the grade 5 test are more difficult than items on the grade 4 test. The item parameters for 100 items (50 per grade) could be estimated with a 3PLM. However, only the information about the 15 common items would be used to link the two tests. Given the IRT property of parameter invariance, the discrimination, difficulty, and guessing parameter of each common item ($i = 1, \ldots, 15$) on the grade 5 test (α_{i5}, β_{i5}, γ_{i5}) will have the

following relationship with the corresponding parameters for the same items on the grade 4 test (α_{i4}, β_{i4}, γ_{i4})

$$\alpha_{i4} = \frac{\alpha_{i5}}{A}$$
$$\beta_{i4} = A\beta_{i5} + B$$
$$\gamma_{i4} = \gamma_{i5}$$

Again, *A* and *B* are the linking constants used for the linear transformation. Because of the relationship between the item parameters and θ, only one set of linking constants is needed to transform all of the item parameters to the grade 4 scale. Conceptually, the *A* constant adjusts the discrimination of the items and the *B* constant adjusts the difficulty. Note that there is no change to the guessing parameters. While the equations above capture the theoretical relationship between the common items on the two tests, the linking constants are still unknown. Various methods have been proposed to estimate these constants, but the Stocking- Lord method (Stocking & Lord, 1983), which minimizes the difference between test characteristic curves represented by the common items between grades, is most commonly used.

When there is a need to place scores from more than two tests on a common scale, the linear transformation presented above can be extended using a process known as "chain linking." That is, if there are common items between grades 4 and 5, and a separate set of common items between grades 5 and 6, linking constants can first be estimated for each grade pair (C_{45} and C_{56}). The grade 5 scores would be transformed to the grade 4 scale—the base grade—using the C_{45} constants, and the grade 6 scores would be transformed to the grade 4 scale by first transforming them to the grade 5 scale using the C_{56} constants, and then the C_{45} constants.

In contrast to the separate linking approach described above, under concurrent linking, all item parameters for all grades are estimated in one step during which different underlying

population ability distributions are specified for each group of students taking the tests across grades (Bock & Zimowski, 1997). With the differences in ability for the two populations of test-takers taken into account directly, the parameters of items common to both groups serve to anchor the score scale. If this approach were used in the hypothetical example above, 85 sets of item parameters would be estimated: 70 for the unique items and 15 for the common items. All 85 items are automatically calibrated to be on the same scale with the 15 common items providing the link between the two tests.

Choosing an Ability Estimator

Once item parameters have been estimated (whether separately or concurrently), a specific scale score can then be estimated for each individual. The two most common approaches for accomplishing this are maximum likelihood (ML) estimation and expected *a posteriori* (EAP)[4] estimation (Bock & Aitkin, 1981). With the ML approach, the joint probability of an examinee's response pattern is maximized to determine the most likely ability level. With the EAP approach, the joint probability distribution is weighted by a set of quadrature points—typically associated with a normal distribution—to provide an estimate of ability. The key tradeoff between these two methods is one of efficiency versus bias. ML estimates are asymptotically consistent, but they can be inefficient for examinees with ability near the tails of the distribution. EAP estimates are biased, but they are easily calculated, and minimize measurement error Bock and Mislevy (1982, p. 439). The estimation of a student-specific scale scores by grade is the last step before it becomes possible to make comparable interpretations of longitudinal growth trajectories.

A Brief Review of the Literature on IRT-based Vertical Scaling

Research on the use of IRT to create a vertical score scale (sometimes referred to as a "developmental" scale) dates back to the 1970s. Much of the earliest research focused on the use of the 1PLM to link tests that differed in difficulty (Divgi, 1981; Gustaffson, 1979; Holmes, 1982; Loyd & Hoover, 1980; Slinde & Linn, 1978; 1979a; 1979b; Wright, 1977); subsequent research compared the use of both the 1PLM and the 3PLM to more traditional approaches (i.e., equipercentile methods, Thurstone scaling) (Kolen, 1981; Marco, Petersen, & Stewart, 1983; Petersen, Cook & Stocking, 1983; Skaggs & Lissitz, 1986). By the mid 1980s, these comparisons had led to one rather puzzling finding: IRT-based vertical scales indicated that within-grade variability in test performance decreased over time, a phenomenon described as "scale shrinkage." This was in contrast to the findings from traditional scaling approaches, which generally showed the opposite pattern. Camilli (1987), Camilli, Yamamato & Wang(1993), and Williams, Pommerich & Thissen (1998) speculated that the decreasing variability could be explained by problems associated with IRT ability estimation for low and high-scoring students across grade level tests. The latter may have been due to the extensive use of the IRT software LOGIST (Wingersky, Barton & Lord, 1982) and BICAL (Wright & Mead, 1978). Both programs depend upon joint maximum likelihood (JML) estimation of item and person parameters, and the use of JML has been shown to lead to bias these estimates under certain conditions. (For details, see Baker & Kim, 2004.) Kolen (2006), notes that among studies using marginal maximum likelihood (MML) estimation (Bock & Aitken, 1981), there has been no consistent pattern of decreasing within-grade variability. When a pattern of decreasing

variability does occur, one might reasonably argue on the basis of research by Yen (1985) that a violation of the IRT assumption of unidimensionality is a likely culprit.

Our review of this research literature has led us to focus on three factors which we will be varying in the process of creating a vertical scale: (1) the specification of the underlying IRT model, (2) whether linking is done separately or concurrently, and (3) whether proficiency estimation is based up on a ML or EAP approach. Below we explain the rationale behind our manipulation of each factor, and to what extent we can predict the marginal impact each factor should have on subsequent scale interpretations.

What IRT Model Should be Specified?

The decision about which IRT model to use when scaling a standardized test may be made for statistical, pragmatic and/or even philosophical reasons[5]. From a statistical perspective, more complex models such as the 3PLM and GPCM will always fit the data better than more parsimonious models such as the 1PLM and PCM, and will provide for more precise estimates of examinee proficiency. On the other hand, if there is an interest in developing a score scale with interval properties, then at least in theory, this is somewhat more plausible using a combination of the 1PLM and GPCM. Finally, more parsimonious models tend to lend themselves to more transparent interpretations. The 3PLM and GPCM weight items relative to their discrimination parameters, whereas item weight is constant under the 1PLM and PCM. As such, provided there is no missing data, the rank orders of student raw scores and 1PLM/PCM scores will be the same. The same is generally not true for the 3PLM and GPCM[6].

When IRT models are used to craft a vertical scale, there is evidence to suggest that the

more complex specification is preferable to the parsimonious one. In their review of the literature, Skaggs and Lissitz (1986) concluded that use of the 1PLM has been shown to produce "inadequate results" and counseled practitioners towards cautious use of the 3PLM instead. The main problem found with use of the 1PLM in vertical scaling appears to have been a failure to account for guessing when the scale score of examinees with low proficiency are linked on the basis of items that are very difficult for them to answer.

We have chosen to create vertical scales using the 1PLM and PCM as a contrast to the use of the 3PLM and GPCM for two reasons in particular. First, in practice, many large-scale assessments are in fact based on the use of the 1PLM and PCM to create a vertical scale. Hence, it is important to demonstrate the impact this choice will have on VAM estimates relative to the use of a more complex IRT model. Second, many of the studies that have been pessimistic about the use of the 1PLM for purposes of vertical scaling were either using tests with items that did not fit the model (Slinde & Linn, 1978; Divgi, 1981) or were estimating both item parameters and student proficiency using JML. In the present study, we are using a test in which the fit of the items to the 1PLM appears relatively good, and all parameter estimation is based on MML techniques. One consequence that we can anticipate when comparing grade by grade growth from the 1PLM/PCM to the 3PLM/GPCM is that with the latter the score scale will be expanded relative to the former, for reasons we describe in the next section.

Separate or Concurrent Linking?

Separate and concurrent linking approaches each have strengths and weaknesses. The separate approach is easy to implement, but because it is unlikely that linking constants are

estimated without error, to some extent this additional error will contaminate the transformed scale. As Kim, Lee, & Kim (2008) have shown, this will particularly be the case as the transformed grade departs further from the base grade. In contrast, with the concurrent approach only one model must be specified to estimate all the parameters and create the vertical scale. In this regard there is no comparable source for "linking error," unless one were to conceptualize such error with respect to the choice of common items in the underlying linking design (Michaelides & Haertel, 2004). In practice, testing companies have been more likely to build and maintain a vertical scale using a separate linking approach. There are at least two reasons for this. First, because testing companies typically employ large item banks, it is often unfeasible to estimate parameters for all items simultaneously when new items are added. Second, and probably more importantly, when the definition of the measured construct changes across grades, concurrent estimation can introduce bias throughout the entire scale because the assumption of unidimensionality has been violated, whereas separate linking may mitigate such bias by relying only on pairwise linking across grades (Béguin & Hanson, 2001; Béguin, Hanson, & Glas, 2000).

The debate over the use of separate versus concurrent approaches is relatively new to the research literature, and at present there is no consensus—either theoretically or empirically—as to which should be preferred (Ito, Sykes & Yao, 2008; Hanson & Béguin, 2002; Kim & Cohen, 1998). In the present study, we develop vertical scales using a separate approach, and a modification of a purely concurrent approach. In most applications of IRT-based vertical scaling, a cross-section of examinees in a single year are tested across a vertical grade span, and this becomes the basis for establishing a vertical link across the tests. Instead, as we describe in more detail in the next section, our data comes from two longitudinal cohorts of students—the same

students are included in the dataset on multiple occasions. This created violations of the IRT local independence assumption that we were unable to resolve with the multigroup IRT software at our disposal. For this reason, instead of using a purely concurrent approach, we used a hybrid method similar to the one taken by Karkee, Lewis, Hoskens, Yao & Haug (2003) in which item parameters for the within-grade/across-year assessments were estimated concurrently, and then separate calibration was used to create the across-grade links. If scales differ by calibration approach, the most likely explanation is linking error, multidimensionality, or some combination of the two.

Ability Estimation: ML or EAP?

The choice of ability estimator is also relatively new to the research literature, but it is a factor flagged by both Kolen & Brennan (2004) and Kolen (2006) as an important distinguishing feature among vertical scales. For states scaling their tests with the 1PLM and PCM with the software Winsteps (Linacre & Wright, 1998), a choice of ML estimation is implicit. For states scaling their tests with the 3PLM and GPCM, a default choice of EAP estimation is usually implicit. The impact of this choice will have a predictable impact on growth interpretations: the use of EAP estimates will contract the scale relative to the use of ML estimates. Because EAP estimates are shrunken to a population mean, they will be less variable than ML estimates. On the other hand, when aggregated across students and schools at the state level, both EAP and ML estimates should have the same population mean.

Methods

Data

The data we use in this study are longitudinal item responses from the Colorado Students Assessment Program (CSAP) reading test. The CSAP vertical scale is based on a common item nonequivalent groups linking design that was established by the state's test contractor, CTB/McGraw-Hill, in 2001. The vertical scale was created by scaling each grade-specific test from 3-10 with the 3PLM and GPCM, linking the tests together using a separate approach (with grade 7 as the base grade of the scale), and then producing ML ability estimates for student-level scores. Since its creation in 2001, parallel test forms have been administered at each grade. Item parameters and ability estimates from subsequent tests are horizontally equated so that they can be compared relative to the 2001 scale (i.e., the base scale). To maintain the vertical scale within this linking framework, there is no single set of common items across years in each grade; rather, different sets of items are shared across different years, and there are a limited number of across-grade common items in any given year.

The vertical scales created in the present study derive from longitudinal 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. Our vertical scaling design involved the linking of tests for adjacent grades in the same year, as well as the linking of tests for the same grade in adjacent years. This design is summarized in Table 1 below. The CSAP reading

tests used to create different vertical scales contained a mix of multiple-choice (MC) and constructed-response (CR) items. In grade 3 the test consisted of 34 MC items and 7 CR items; in grades 4-7 the respective numbers were about 70 MC items and 14 CR items. The number of common MC and CR items across adjacent grades or years ranged from 7 to 20 and 0 to 4 respectively. Each of the eight grade by year combinations (grades 3 to 7 in 2003-2006) used for the analysis included an average of 55,681 students enrolled in 1,379 unique public schools. Roughly 64% of the students self-identified as white, 26% as Hispanic, 6.2% as black, 3% as Asian/Pacific Islander, and 1.3% as Native American.

Insert Table 1 here

While the design we use to create our vertical scales—common item nonequivalent groups—is consistent with the design underlying the CSAP operational scale, it differs with respect to (a) our choice of base year and grade (2003 and 3 respectively), and (b) the number of common items available to link tests across grades or across years. As a rule of thumb, Kolen & Brennan (2004, p. 271) have recommended that the number of common items available for test linking equal roughly 20% of the total number of items on any single test form. In our design there is one link that clearly falls short of this rule of thumb: for the grade 6 tests between 2005 and 2006 there are just 7 common items (about 10% of the total items). Given this limitation in our design, inferences about scale score growth that involve grade 6 test scores should be made with caution.

Creating Vertical Scales

We created eight vertical scales that differ with regard to three factors: (1) the IRT model used to estimate item parameters, (2) the linking method used to place the parameters from different grades onto a common scale, and (3) the method used to estimate student-level scale scores. Table 2 provides an overview of the eight scales that result from the combination of these variables. The operational CSAP vertical scale for reading is currently based on the combination of factors represented in cell number 1: 3PLM/GPCM, separate linking, and EAP score estimation. However, as discussed in our review of the literature, the scales represented by cells 2-7 would be defensible alternatives.

Insert Table 2 here

The two different linking approaches taken in this study are illustrated graphically in Figures 2 and 3. In Figure 2, each oval represents a separate linking of tests across grades in the same year (vertical direction), or across years in the same grade (horizontal direction). Under the hybrid approach illustrated in Figure 3, we separately linked tests across grades in the same year (ovals), but concurrently linked tests across two years in the same grade (rectangles). When using the separate approach, we first estimated item parameters for each grade estimated independently. Next, using the Stocking-Lord method (Stocking & Lord, 1983), we estimated linking constants for the various within-year/across-grade and within-grade/across-year pairs

illustrated in Figure 2. We used these constants to transform both item parameters and estimates of latent proficiency, θ, onto the grade 3 scale using chain linking. Linking constants and the chain linking transformation were computed using the R package plink (Weeks, 2007). Under the hybrid approach, we first estimated item parameters for each of the within-grade/across-year assessments concurrently. Next, we used the Stocking-Lord method to estimate linking constants for the across-grade linkages shown in Figure 3. The item parameters and proficiency estimates for these scales were then placed onto the grade 3 scale using chain linking. All of the item parameters were estimated using an MML approach in the IRT Command Language program (ICL; Hanson, 2002). We report scores for each vertical scale in logit units.

Insert Figures 1 and 2 here

Value-Added Model: The Layered Model

To estimate value-added effects for schools on a particular grade of students we specified a constrained version of the general value-added model first described by McCaffrey, et al.. (2004), and then later named the *variable persistence model* by Lockwood, et al.. (2007). This constrained model is equivalent to a single cohort longitudinal version of the layered model popularized by William Sanders and colleagues (cf., Sanders et al., 1997). Note that because the model only considers longitudinal data for a single cohort of students, in this context a "school effect" and a "grade effect" are the same thing. The variable persistence model takes the following form:

$$Y_{it} = \mu_t + \sum_{t^* \le t} \alpha_{tt^*} \grave{e}_{t^*} + \varepsilon_{it}. \qquad (3)$$

In equation 3, Y_{it} represents the CSAP reading test score for student *i* in year *t*, *t* = 1, …, *T*, and the parameter μ_t denotes the grand test score mean for a given year. The vector $\grave{e}_{t^*}$ represents the collection of school effects[7] for each year, and the parameter α_{tt^*} captures the persistence of the school effects $\grave{e}_{t^*}$ in year *t* (given that $t^* \le t$). Finally, ε_{it} represents the test score residual associated with student *i* in year *t*. Both $\grave{e}_{t^*}$ and ε_{it} are assumed to be independent latent random variables, where $\varepsilon_{it} \sim N(\mathbf{0}, \mathbf{Ó})$ and $\grave{e}_{t^*} \sim N(0, \tau)$. To be consistent with the assumptions of the layered model, equation 3 is constrained such that all persistence parameters are set equal to 1 ($\alpha_{tt^*} \equiv 1$ for all $t^* \le t$)[8].

Applying the model above to each of the eight vertical scales we created for the time period from 2003 to 2006 yields the following system of equations

$$\begin{aligned}
Y_{i03} &= \mu_{03} + \grave{e}_{03} + \varepsilon_{i03} \\
Y_{i04} &= \mu_{04} + \grave{e}_{03} + \grave{e}_{04} + \varepsilon_{i04} \\
Y_{i05} &= \mu_{05} + \grave{e}_{03} + \grave{e}_{04} + \grave{e}_{05} + \varepsilon_{i05} \\
Y_{i06} &= \mu_{06} + \grave{e}_{03} + \grave{e}_{04} + \grave{e}_{05} + \grave{e}_{06} + \varepsilon_{i06}.
\end{aligned} \qquad (3)$$

We note the following about the school-level parameters in these equations. First, the parameter vectors $\{\grave{e}_{04}, \grave{e}_{05}, \grave{e}_{06}\}$ 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 $\grave{e}_{03}$, which captures pre-existing differences in school status as of grade 3. Second, while the model above can be easily extended to allow for multivariate test outcomes (typical of applications of the layered model), 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.., 2007a,

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 Lockwood (2006) and described by Lockwood et al. (2007a). 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, and the standard deviation of this distribution as an estimate of the uncertainty. 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.

Results

Comparing Vertical Scales

The means and standard deviations (SDs) for each of the eight vertical scales we created are summarized in Figures 4 and 5 for the first of our two longitudinal student cohorts[9]: those students who were in grade 3 in 2003 and grade 6 in 2006. For each of these two statistics, there are three "main effects" of interest:

1. The difference between IRT models used to estimate the item parameters (1PLM/PCM vs. 3PLM/GPCM). All scales that involved the 1PLM/PCM

combination are denoted graphically by lightly shaded lines, while all scales that involved the 3PLM/GPCM are denoted by darkly shaded lines

2. The difference between approaches used to link the vertical scales (separate vs. hybrid). All scales that were created using the separate approach exclusively are denoted graphically by solid lines, and the scales created using the hybrid approach are denoted with dotted lines.
3. The difference between approaches used to estimate student-level scale scores (EAP vs. ML). All scales that were created with EAP estimation are denoted graphically with X markers, and the scales created with ML estimation are denoted with O markers.

Insert Figures 4 and 5 here

There are clear patterns in the changes in means and SDs for the different vertical scale across grades. We start by examining the different trends in growth by scale shown in Figure 4. For each scale, the growth in score means from grade to grade appears somewhat nonlinear, and decelerating over time. This is consistent with previous findings in the literature (Kolen, 2006, p. 178). The apparent magnitude of growth along the logit scale differs substantially as a function of the underlying IRT model and linking approach. Scales created using the 3PLM/GPCM and separate linking combination give the impression of the most growth; scales created using the 1PLM/PCM and either separate or hybrid linking give the impression of the least growth. This was to be expected. The use of the 3PLM/GPCM stretches the score scale because distinct scores are computed for each unique item response pattern. This is in contrast to

the 1PLM/PCM, where an examinee's total score serves as a sufficient statistic for his or her scale score; distinct scale scores are computed for each raw score. What was not necessarily expected was that when 3PLM/GPCM estimates are linked using the hybrid approach rather than separate approach, the observed growth trajectory is shifted downward.

Figure 5 plots the within-grade variability over time for each scale. The first thing to note is that the pattern from grade to grade is inconsistent. From grade 3 to 4, score variability decreases significantly across all but one scale, whereas from grade 5 to 6, it stays roughly constant or increases slightly across all eight scales. Between grades 4 and 5, variability increases for the two scales based on the 3PLM/GPCM that use hybrid linking with ML or EAP estimation, but decreases for all other scales. In general, using the 3PLM/GPCM produces scales with greater variability than the 1PLM/PCM, the use of ML estimation increases the variability in a scale relative to EAP estimation, and the use of hybrid linking appears to decreases scale variability relative to separate calibration. So while the choice of IRT model clearly has the largest impact on scale variability, the choice of linking and ability estimation approach still have a significant impact, especially when combined with use of the 3PLM/GPCM.

One message conveyed by viewing Figure 4 in conjunction with Figure 5 is that interpretations of grade by grade growth along a vertical score scale can be misleading unless one also takes into account the associated variability of the scale. This is illustrated in Figure 6, which for illustrative purposes contrasts two extremes: a vertical scale based on the 3PLM/GPCM, separate linking, and ML estimation (the current approach taken in Colorado); and a vertical scale based on the 1PLM/PCM, hybrid linking, and EAP estimation. In terms of the three decisions used to establish each vertical scale in each case, the former scale maximizes score variability while the latter scale minimizes it.

Insert Figure 6 here

Year to year growth along a given scale can be compared while adjusting for the variability of the underlying scale by standardizing the mean differences. Yen (1986) defined an effect size statistic for these purposes as

Effect Size = ,

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. All else held constant, it will always be the case that effect sizes based on EAP proficiency estimates will be larger than those based on ML proficiency estimates.

Insert Figures 7 and 8 here

The effect size estimates that correspond to the growth from grades 3 to 4, 4 to 5 and 5 to 6 are shown in Figures 7 and 8 for the four 1PLM/PCM scales and 3PLM/GPCM scales respectively. In general, these effect size patterns are fairly similar, however for each across grade comparison, the effect sizes based on two different scales can differ by as much as 10 to 20 percent of the average SD across grades. We note that though there are considerable differences between the effect sizes for growth from grade 3 to 4 as a function of ability estimation

approach, these differences are negligible by the time effect sizes are compared for growth from grade 5 to 6. We have no explanation for this finding at the present time. Finally, we note that the use of the 3PLM/GPCM does not always result in the largest effect sizes. When the latter is used in conjunction with ML estimation, the resulting growth can be anywhere from 5 to 10% smaller in effect size units relative to a vertical scale based upon the 1PLM/PCM and EAP estimation.

Comparing Value-Added School Effect Estimates from the Layered Model

The layered model specified by equation 3 was used to estimate school effects for each of the eight sets of longitudinal scale scores described above. Below we present correlations of estimated school effects across scales as well as discrepancies in resulting classifications of schools as "effective" or "ineffective" [10]. The correlations, by grade, between the school effects estimated using the layered model are all very strong and positive, ranging from a low of .79 to a high of .99 with a mean of .95. In other words, although the various scales differ with regard to growth in an absolute sense, they convey a similar message about the ordering of school effects. In Table 3 we compare, for each grade, the number and percent of schools that would be classified as above average, average and below average in terms of the value they add to student achievement. A school is classified as above average if its value-added effect estimate is more than one posterior standard deviation above zero and "below average" if its estimated effect is more than one standard deviation below zero. We use this particular approach to classify schools to be consistent with the approach taken by McCaffrey et al (2004) and Lockwood et al (2007a) to classify teachers when using the same VAM considered here.

Insert Table 3 here

The results in Table 3 support the following three conclusions:

1. The greatest discrepancy in the percentage of schools classified as "above average" across scales is 7, 6, and 5 percentage points for grades 4, 5 and 6, respectively. The corresponding discrepancies for schools identified as "below average" are 5, 2 and 6 percentage points.
2. More schools can be reliably classified as above or below average using scales created with the 3PLM/GPCM than scales created with the 1PLM/PCM..
3. More schools are reliably classified as being above or below average when the scales are based upon EAP rather than ML estimates of student achievement.

Within each combination of IRT model and estimation approach, the choice of linking approach (separate or hybrid) makes little difference in the percent of schools classified as above or below average. This is as one would expect for the type of value-added model we have specified here, an issue we will return to in the next section. One conclusion to be drawn from these results is that none of the three variables used to create the vertical scales (IRT model, calibration approach, estimation approach) appear to have a large independent impact on the estimated school effects under the layered model. However, particular combinations of these three variables can lead to significant differences in the precision of the numbers of schools classified as above or below average in their effectiveness. To illustrate this, Tables 4 through 6 compare the number of schools that can be reliably classified as "above average" (+), "average" (0) or "below average" (-) on the basis of the value they appear to have added to student reading

performance in grades 4 through 6. The rows and columns represent school classifications under the "Separate 3PL EAP" and "Hybrid 1PL MLE" scales respectively. Of interest are the numbers of schools in the off-diagonals; whereas one vertical scale would identify such a school as "effective", a different scale would identify it as "ineffective". Of the grade 5 effects, a total of 82 schools (out of 950) would be classified as ineffective under one scale, but average under the other; another 73 would be classified as effective under one scale, but average under the other[11]. If sanctions or rewards are attached to these classifications, the choice of scaling approach can clearly have important ramifications.

Insert Tables 4-6 here

Discussion

Using longitudinal growth in student achievement as the basis for evaluating school performance in an accountability system is a methodological approach that is gaining steam. Due to the simple fact that value-added models use students as their own controls, such an approach would appear to address the well-understood "Beverly Hills" problem that confounds accountability decisions associated with NCLB that are based solely on school-level status: the schools making adequate yearly progress tend to be located in wealthy communities. In contrast, our estimates of school-level effects for grades 4 through 6 are uncorrelated with school-level proxies for poverty (percent of students eligible for free and reduced lunch services) as well as levels of test score performance. In other words, after applying the layered model to Colorado data, wealthy and high-achieving schools are no more likely to have positive school effects than

are poor and low-achieving schools. In this sense value-added modeling approaches provide an appealing alternative and/or complement to accountability systems based solely on criterion-based measures of status. Nonetheless, such approaches come at the cost of great statistical complexity and potentially misguided causal inferences (Braun, 2005; Briggs & Wiley, 2008; Raudenbush, 2004; Rubin, Stuart & Zanatto, 2004).

One key assumption that can easily be overlooked is the manner in which student achievement is being measured and vertically scaled. In this study we have conducted an empirical sensitivity analysis by (a) gathering longitudinal item response for two cohorts of students who were administered Colorado's CSAP reading test between 2003 and 2006, (b) creating eight defensible vertical scales with this data, and (c) using the resulting scales as the outcome variable in a commonly used value-added model At the outset of this paper we posed two research questions. We now summarize our findings with respect to each question.

What is the sensitivity of a longitudinal score scale to the way the test scores have been vertically scaled?

The longitudinal score scales that are established using IRT-based approaches have no absolute interpretation. Depending upon the underlying IRT model, the linking approach and the estimation approach that are taken, the score scale can be, in effect, stretched or compressed. It follows from this that if one only interprets mean growth over time without taking the variability of the scale into consideration, then a longitudinal score scale is very sensitive to the way a vertical scale has been created. When the scale is interpreted in effect size units such that information about mean changes and scale variability are combined into a single statistic, growth

patterns are more similar, but there are still some substantive differences across scales—as much as 20% of a standard deviation. In other words, even when considering the same item responses on the same tests from the same populations of students, absolute interpretations of growth in reading achievement can be influenced by the way the underlying scale has been established. In state educational accountability systems with tests that have been vertically scaled, scales scores are ultimately converted into discrete performance categories (i.e., "proficient", "advanced", etc.) by establishing cut-points on the scale through the process of standard-setting. It is an open question whether this process can successfully give criterion-based meaning to these cut-points that do not depend on the properties of the underlying vertical scale.

What impact do different vertical scaling approaches have on estimates of value-added school effects?

We estimated grade 4, 5 and 6 value-added school effect estimates as a function of our eight vertical scales and correlated the results. In general, the correlations were very strong and positive. This is not surprising because value-added estimates are inherently norm-referenced; so long as year to year changes in the score scale impact schools in the same way, the ordering of value-added residuals will be unaffected. On the other hand, we found that the numbers of schools that could be reliably classified as effective, average or ineffective was somewhat sensitive to the choice of the underlying vertical scale. When VAMs are being used for the purposes of high-stakes accountability decisions, this sensitivity might be considered problematic.

The VAM we specified for our data is a version of the layered model that has become popular as a component of state educational accountability systems in Tennessee, Ohio and

Pennsylvania. Because the VAM residuals estimated in this model are based in part upon the prediction of present test performance on the basis of past test performance, there is no requirement that the tests be on the same score scale. (A decision must be made about the appropriate score scale, but not about the scales vertical properties across grades.) Hence we did not expect estimated school effects to be sensitive to differences in specific approaches (i.e., separate, hybrid) taken to link test scores across grades, and our results bear out this expectation. For the sorts of VAM specifications, it is only the choice of score scale creation within each grade that can make a difference to school effect estimates. On the other hand, the need for tests with a vertical score scale across grades *is* a requirement for VAM specifications in which year to year growth has either a parametric function (i.e., the hierarchical linear models popularized by Raudenbush & Bryk, 2002), or when difference scores are used as the dependent variable in an econometric regression model. In such cases, while the specific choice of linking approach may not lead to dramatically different value-added effect estimates, the choice not to link the tests at all almost certainly will.

Limitations

Earlier we noted an important limitation to this study related to the design of the CSAP reading assessment. Namely, the common item nonequivalent groups design of our data was not a variable we were able to manipulate. Because the growth along any vertical scale will depend upon the common items chosen to overlap between grades, the extent to which our results are sensitive to the common items that were pre-established for the CSAP is unknown. In addition,

our analysis has only considered differences among vertical scales in the subject of reading. Whether similar growth patterns would be found for math and science is unclear.

There are many different ways to specify value-added models beyond our choice of the layered model. Hence it might be worthwhile to more fully extend our comparisons to other VAM specifications, and to also consider the impact when no attempt is made to convert the raw scores on a test (i.e., number correct) into a scale using IRT. Finally, there is reason to suspect that VAM estimates will be most sensitive to nonlinear manipulations of the underlying score scale (Ballou, this issue). Yet because the both the 1PLM/PCM and 3PLM/GPCM had showed good fit to the CSAP tests used in this study, each of the eight vertical scales that were created were effectively linear transformations of one another. As a result the findings here may well understate the potential sensitivity of value-added estimates to the way test scores are scaled, vertically or otherwise.

Future Directions

The contribution of the present study is to demonstrate that the choice of vertical scaling approach can have significant impact on the precision of school-level classifications within an educational accountability system. In this sense our findings are similar in spirit to those of McCaffrey et al. (2004) and Lockwood et al. (2007a) who showed (among other things) that the precision of teacher-level classifications depends on the way that that the persistence of teacher effects is parameterized. It is important to note that if the choice of vertical scale only affects the precision of value-added estimates, this in and of itself may not raise serious red flags about the use of value-added models for school or teacher accountability. This is because there are ways

to compensate for a loss of precision that are tied to the manner in which vertical scales are developed or how teacher/school effects are estimated. For example, value-added estimates could be aggregated over several longitudinal cohorts or over several test subjects, or a variable persistence model could be specified instead of a complete persistence model.

While the choice of scale appears to have a significant impact on the precision of value-added estimates, these estimates remain strongly correlated across scales. This would appear to suggest that norm-referenced orderings of schools are unlikely to depend upon the technical decisions made in creating a vertical scale. However, it is important to note that we have made the questionable (though commonplace) assumption that the construct of reading comprehension maintains a unidimensional interpretation over a five year grade span. It has previously been established that when tests measuring multiple dimensions are modeled using unidimensional methods, ability estimates will be biased (Ackerman, 1992; Beguin, Hanson, & Glas, 2000). This problem is further exacerbated when the dimensional structure changes from test to test over time—when there is what Martineau (2006) calls "construct shift." If construct shift over grades is occurring but is not modeled explicitly, the scores along a unidimensional vertical scale will be biased. This may be the explanation for the finding of decelerating growth across most vertical scales that span multiple school grades. If these scores are biased, it follows that value-added estimates will also be biased. Hence the fact that all the value-added school effects based on the 8 vertical scales in our study are strongly correlated could be misleading if they each contain a substantial amount of bias because they ignore multidimensionality. If value-added models are being applied to large-scale assessments that have substantive multidimensional interpretations, this is cause for concern. Lockwood et al. (2007b) showed that value-added effects are much more sensitive to the dimensionality of the outcome being modeled than they

are to the choice of value-added model used to estimate the effects. As such, we plan to tackle this issue with the same data in future research by attempting to create multidimensional vertical scales, and examining the sensitivity to subsequent estimates of both growth and value-added.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Baker, F. B., & Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques* (2nd ed.). New York: Marcel Dekker.

Ballou, D. (in press). Test scaling and value-added measurement. *Education Finance & Policy* (Special Issue)

Ballou, D., Sanders, W. L., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-66.

Béguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000). *Effect of multidimensionality on separate and concurrent estimation in IRT equating.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Béguin, A. A., & Hanson, B. A. (2001). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

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.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *(6)*4, 431-444..

Bock, R.D., and Zimowski, M. (1997) Multi-group IRT. In W.J. Van der Linden and R.K. Hambleton, (Eds.) *Handbook of modern item response theory*. New York: Springer-Verlag.

Braun, H. (2005, September). *Using student progress to evaluate teachers: A primer on value-added models* [Policy Information Perspective]. New Jersey: ETS.

Briggs, D. C., & Weeks, J. P. (2008) The persistence of value-added school effects. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Briggs, D. C., & Wiley, E. (2008). Causes and effects. In *The Future of Test-Based Educational Accountability,* L. Shepard & K. Ryan (eds). Routledge.

Camilli, G. (1987). Scale shrinkage and the estimation of latent distribution parameters. *Journal of Educational Statistics, 13*(3), 227-241.

Camilli, G., Yamamoto, K., & Wang, M. M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*(4), 379-388.

Colorado Department of Education. (2006). *Colorado student assessment program technical report 2006*. Retrieved January 05, 2009 from http://www.cde.state.co.us/cdeassess/documents/reports/2006/Complete_CSAP_2006_Technical_Report_2006_FINAL.pdf

Divgi, D. R. (1981). Model-free evaluation of equating and scaling. *Applied Psychological Measurement, 5*(2), 203-208.

Gustafsson, J.-E. (1979). The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement, 16*(3), 153-158.

Hanson, B.A. (2002) IRT command language. Monterey, CA: Author (Available online at http://www.b-a-h.com/software/irt/icl/index.html)

Hanson, B. A. & Béguin, A. 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* (1), 3-24.

Harris, D. N. (2008). *The policy uses and "policy validity" of value-added and other teacher quality measures*. Paper presented at the National Conference on Value-Added Modeling.

Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educationa1 Measurement, 19*(2), 139-1 47.

Holland, P., & Dorans, N. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: Praeger.

Ito, K., Sykes, R., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. Applied Measurement in Education, 21, 187-206.

Karkee, T., Lewis, D. M., Hoskins, M., Yao, L., & Haug, C. (2003). *Separate versus concurrent calibration methods in vertical scaling.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Kim, S.-H. & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. Applied Psychological Measurement, 22 (2), 131-143.

Kim, J. K., Lee, W.-C., & Kim, D.-I. (2008). *The effect of choosing a base grade on the vertical scale using various IRT calibration methods.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*(1), 1–11.

Kolen, M. J. (2006). Scaling and Norming. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 155-186). Westport, CT: Praeger.

Kolen, M. J. and Brennan, R. L. (2004). *Test Equating, Scaling and Linking*. 2nd Edition. New York: Springer-Verlag.

Lewis, D. M., & Haug, C. A. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education, 18*(1), 11-43.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement, 24*(2), 115-138.

Linacre, J. M., & Wright, B. D. (1998). *A user's guide to Bigsteps/Winsteps*. Chicago, IL: Mesa Press.

Linn, R.L. (1993) Linking results of distinct assessments. *Applied Measurement in Education*. *6*(1), 83-102.

Lissitz, R. W. & Huynh H. (2003). Vertical equating for state assessments: issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation, 8*(10). Retrieved January 5, 2009 from http://PAREonline.net/getvn.asp?v=8&n=10

Lockwood, J. R. (2006). BTEMS: Bayesian teacher effect modeling software. Pittsburgh, PA: Rand Corporation.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007a). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics, 32*(2), 125-150.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007b) The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*. 44(1), 47-68.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179-193.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), New horizons in testing: Latent trait theory and computerized adaptive testing. New York: Academic Press. (Reprinted from the Computerized Adaptive Testing Conference, 1979, April, Minneapolis)

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. Journal of Educational and Behavioral Statistics, 31(1), 35-62.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.

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

Michaelides, M. P., & Haertel, E. (2004). *Sampling of common items: An unrecognized source of error in test equating* (No. CSE Report 636).

Mislevy, R. J. (1992) *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS Policy Information Center.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Oshima, T. C., Davey, T., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement, 37*(4), 357-373.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*(2), 137-156.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?* Paper presented at the William H. Angoff Memorial Lecture Series, Princeton, NJ. Retrieved from January 25, 2005 from http://www.ets.org/Media/Education_Topics/pdf/angoff9.pdf

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M. D., & Martineau, J. A. (2004). *The vertical scaling of science achievement tests*: Research report for the Center for Education and National Research Council.

Rubin, D. Stuart, A., & Zannato, E. (2004). A potential outcome view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103-116.

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

Seltzer, M., Frank, K. & Bryk, A. (1994). The metric matters: The sensitivity of conclusions concerning growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis, 16*(1), 41-49.

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

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*(4), 495-529.

Slinde, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement, 15*(1), 23-35.

Slinde, J. A., & Linn, R. L. (1979a). A Note on Vertical Equating via the Rasch Model for Groups of Quite Different Ability and Tests of Quite Different Difficulty *Journal of Educational Measurement, 16*(3), 159-165.

Slinde, J. A., & Linn, R. L. (1979b). The Rasch model, objective measurement, equating, and robustness. *Applied Psychological Measurement, 3*(4), 437-452.

Spellings, M. (2005, November 18). Secretary Spellings Announces Growth Model Pilot, Address Chief State School Officers' Annual Policy Forum in Richmond. *U.S.*

Department of Education Press Release. Retrieved January 05, 2009 from http://www.ed.gov/news/pressreleases/2005/11/1182005.html.

Spellings, M. (2007, December 08). Secretary Spellings Invites Eligible States to Submit Innovative Models for Expanded Growth Model Pilot. Retrieved January 05, 2009 from http://www.ed.gov/news/pressreleases/2007/12/12072007.html

Stocking, M. L. and Lord, F. M. (1983) Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2), 201-210.

Thissen, D., & Orlando, M. (2001). Item Response Theory for Items Scored in Two Categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Wainer, H., eds. (2001) *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227-253.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernal method of test equating*. New York: Springer-Verlag.

Weeks. J. P (2007). plink: IRT separate calibration linking methods (R package version 0.0-4). http://cran.r-project.org/web/packages/plink/index.html

Williams, V. S., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement, 35*(2), 93-107.

Wilson, M. (2004). *Constructing measures: an item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1984). LOGIST 5.0 version 2.5 users' guide. Princeton, NJ: Educational Testing Service.

Wright, B. D., & Mead, R. J. (1978) BICAL: Calibrating items and scales with the Rasch model, (Research Memorandum No. 23A). Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D (1997) A history of social science measurement. *Educational Measurement: Issues and Practice*. December 1997, 33-45.

Yen, W. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika, 50*(4), 399-410.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. Journal of Educational Measurement, 23(4), 299-325.

Figure 1. 3PLM Item Characteristic Curve

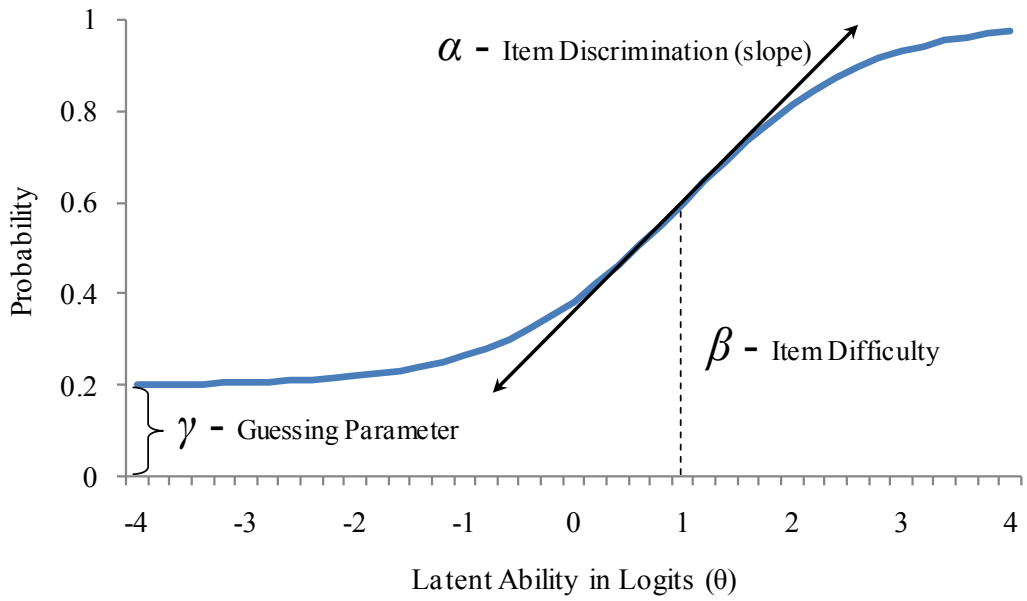


Table 1. Unique and Common Items on CSAP Reading Test by Grade and Year

Grade	Year							
	2003		2004		2005		2006	
3	(34, 7)							
	(13, 3)							
4	(56, 14)	**(15, 3)**	(56, 14)					
			(9, 3)					
5			(56, 14)	**(20, 2)**	(58, 14)			
					(11, 4)			
6					(57, 14)	**(7, 0)**	(57, 14)	
							(10, 4)	
7							(58, 14)	

Note: First value in parenthesis represents number of MC items, second value represents number of CR items. Values in bold represent common items.

Table 2. IRT-Based Vertical Scaling Models

Item Response Model		Linking Approach	
		Separate Calibration	Hybrid Calibration
EAP Scale Scores	3PLM/GPCM	1	2
	1PLM/PCM	3	4
ML Scale Scores	3PLM/GPCM	5	6
	1PLM/PCM	7	8

Figure 2. Separate Calibration Approach

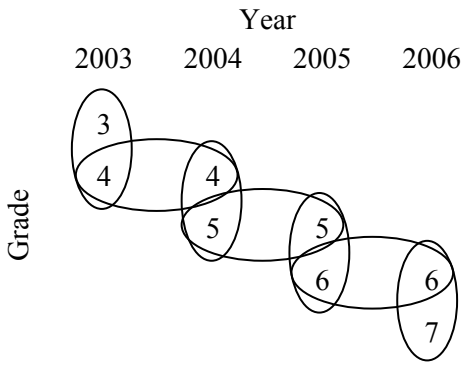


Figure 3. Hybrid Calibration Approach

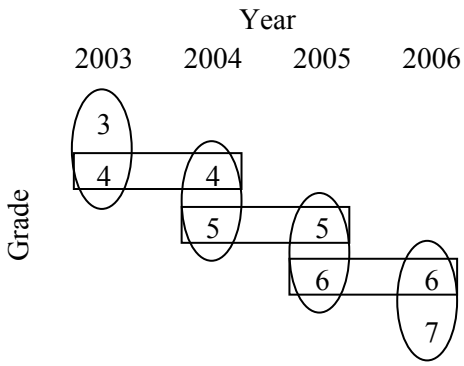
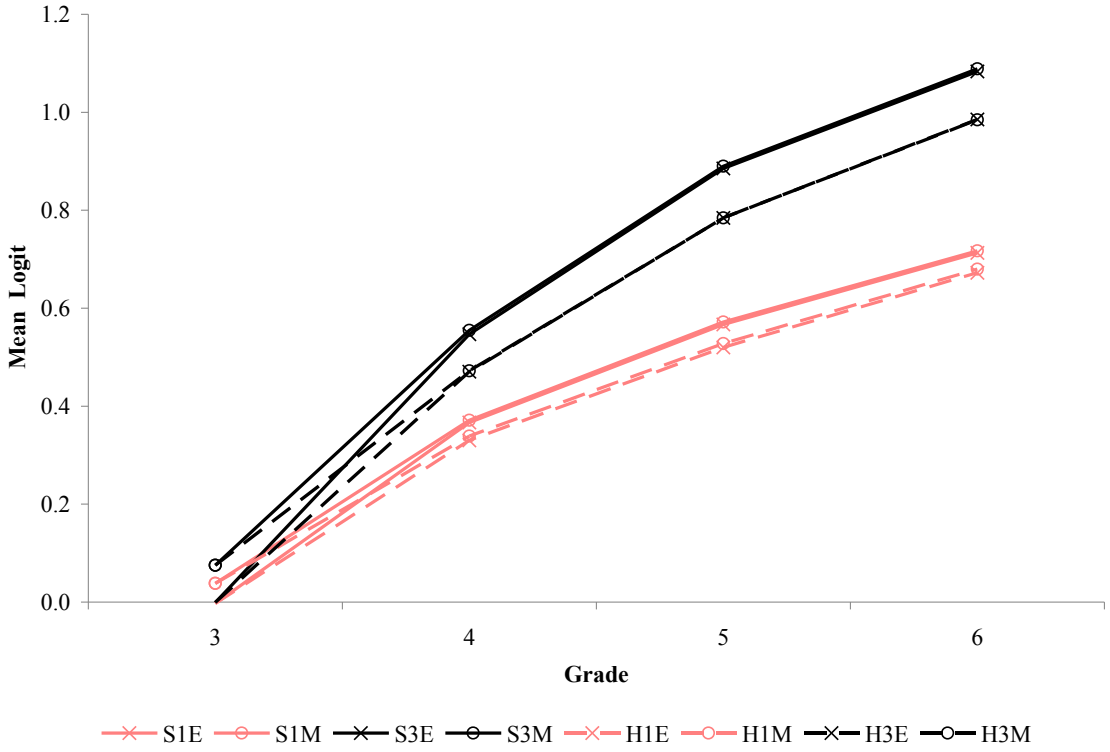
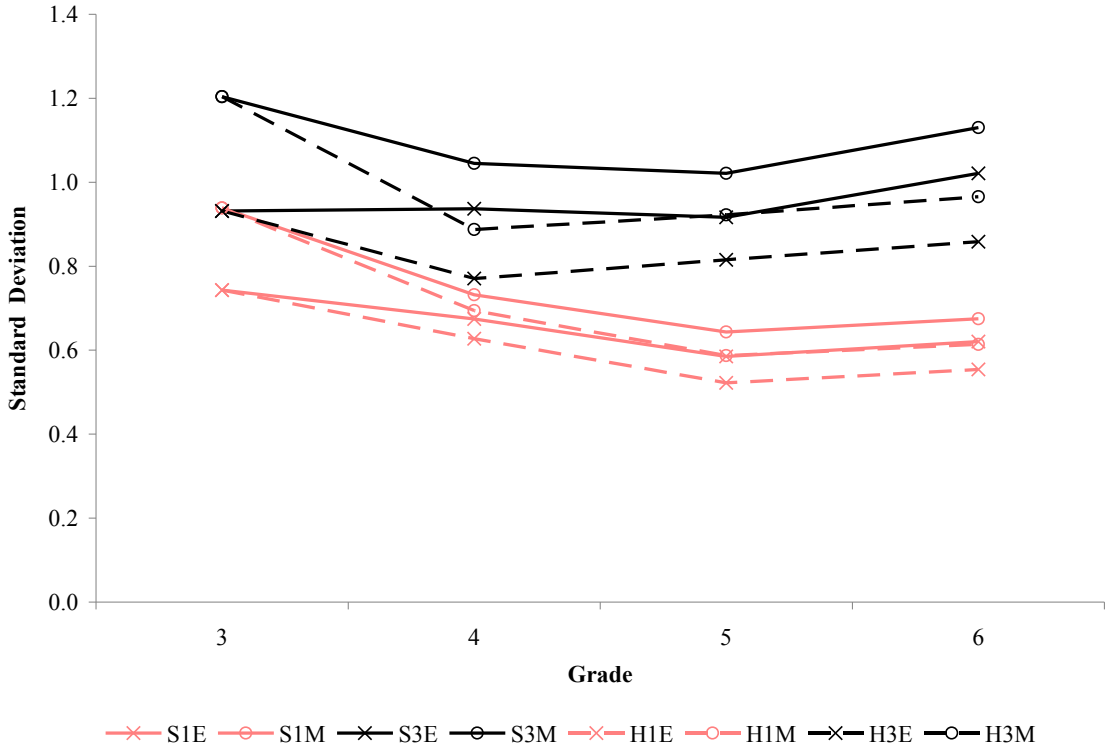


Figure 4. Growth along CSAP Reading Score Scale from 2003 to 2006



Note: In the figure above, "S" = separate linking; "H" = hybrid linking. "1" = use of the 1PLM/PCM; "3" = use of 3PLM/GPCM; "E" = EAP estimation; "M" = ML estimation.

Figure 5. Variability of CSAP Reading Score Scale from 2003 to 2006



Note: In the figure above, "S" = separate linking; "H" = hybrid linking. "1" = use of the 1PLM/PCM; "3" = use of 3PLM/GPCM; "E" = EAP estimation; "M" = ML estimation.

Figure 6. Comparing Extremes in Vertical Scales

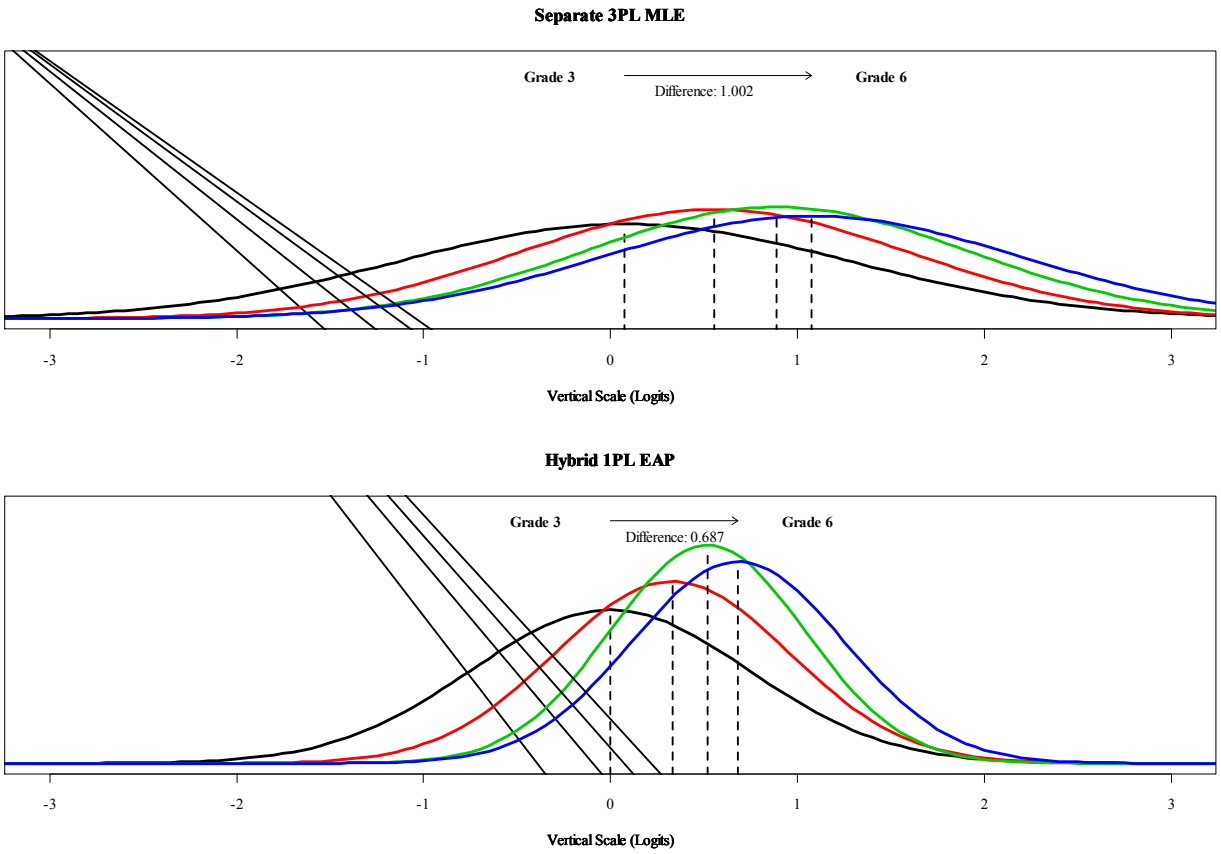
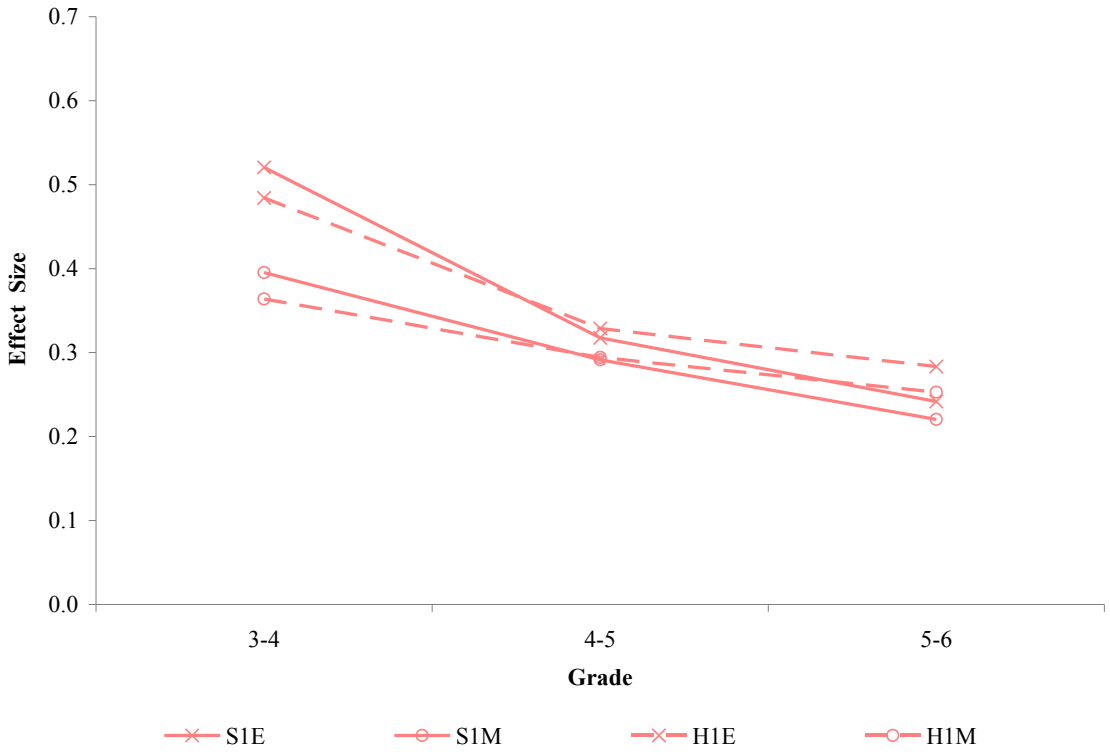
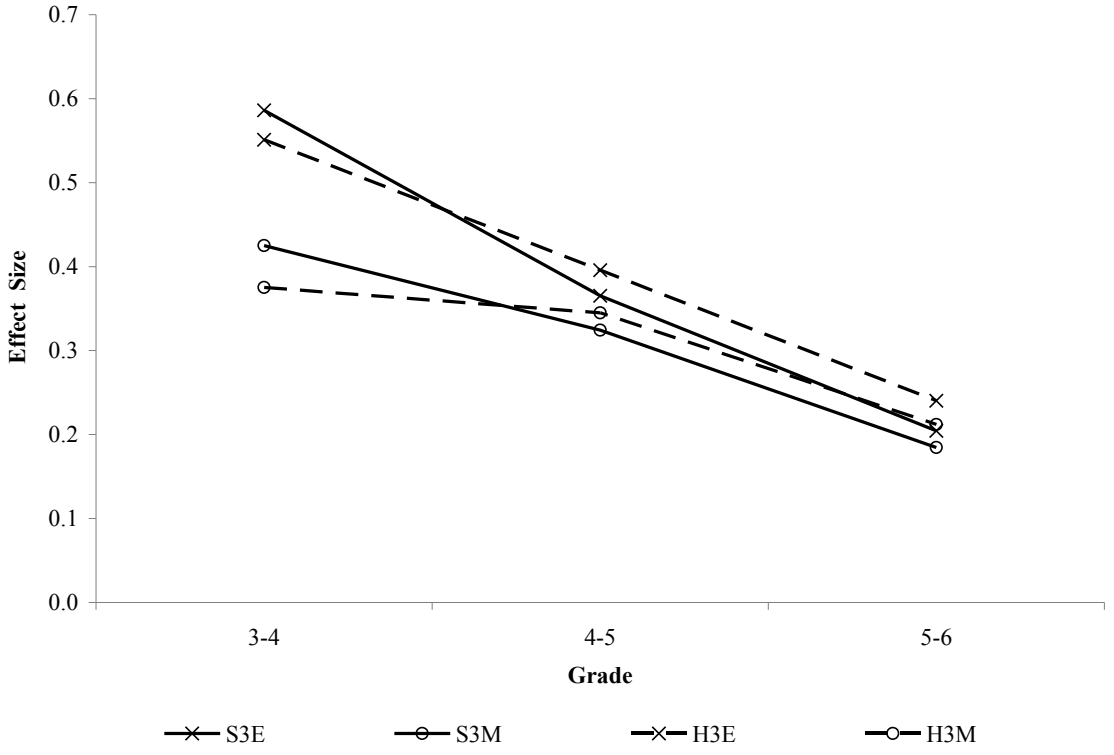


Figure 7. Effect Sizes of Growth for Cohort 1 by 1PLM/PCM Scales



Note: In the figure above, "S" = separate linking; "H" = hybrid linking. "1" = use of the 1PLM/PCM; "3" = use of 3PLM/GPCM; "E" = EAP estimation; "M" = ML estimation.

Figure 8. Effect Sizes of Growth for Cohort 1 by 3PLM/GPCM Scales



Note: In the figure above, "S" = separate linking; "H" = hybrid linking. "1" = use of the 1PLM/PCM; "3" = use of 3PLM/GPCM; "E" = EAP estimation; "M" = ML estimation.

Table 3: Comparison of School Classifications by Underlying Vertical Scale

		Separate Calibration				Hybrid Calibration			
		1PL EAP	1PL MLE	3PL EAP	3PL MLE	1PL EAP	1PL MLE	3PL EAP	3PL MLE
Grade 4 (N=941)	Above Avg.	246 (26)	217 (23)	271 (29)	249 (26)	245 (26)	207 (22)	260 (28)	241 (26)
	Average	512 (54)	576 (61)	479 (51)	532 (57)	518 (55)	591 (63)	498 (53)	550 (58)
	Below Avg.	183 (19)	148 (16)	191 (20)	160 (17)	178 (19)	143 (15)	183 (19)	150 (16)
Grade 5 (N=950)	Above Avg.	221 (23)	221 (23)	242 (25)	233 (25)	208 (22)	213 (22)	263 (28)	255 (27)
	Average	532 (56)	533 (56)	507 (53)	526 (55)	554 (58)	550 (58)	477 (50)	503 (53)
	Below Avg.	197 (21)	196 (21)	201 (21)	191 (20)	188 (20)	187 (20)	210 (22)	192 (20)
Grade 6 (N=640)	Above Avg.	158 (25)	158 (25)	183 (29)	176 (28)	158 (25)	155 (24)	177 (28)	171 (27)
	Average	322 (50)	326 (51)	274 (43)	297 (46)	321 (50)	335 (52)	301 (47)	322 (50)
	Below Avg.	160 (25)	156 (24)	183 (29)	167 (26)	161 (25)	150 (23)	162 (25)	147 (23)

Note for Table 3: School classifications are based upon estimated posterior means and SDs of school effects as specified in the layered model. The category "Above Average" represents a school with an estimated value-added effect that remains above 0 after one posterior SD has been subtracted from its posterior mean. The category "Average" represents a school with an estimated value-added effect that crosses 0 after one posterior SD has been subtracted from or added to its posterior mean. The category "Below Average" represents a school with an estimated value-added effect that remains below 0 after one posterior SD has been added to its posterior mean. Values in parentheses represent column percentages.

Table 4. School Effect Classification by Underlying Vertical Scale-Grade 4

N = 941		Hybrid 1PL MLE		
		+	0	-
Separate 3PL EAP	+	188	82	1
	0	19	422	38
	-	0	87	104

Table 5. School Effect Classification by Underlying Vertical Scale-Grade 5

N = 950		Hybrid 1PL MLE		
		+	0	-
Separate 3PL EAP	+	191	51	0
	0	22	451	34
	-	0	48	153

Table 6. School Effect Classification by Underlying Vertical Scale-Grade 6

N = 640		Hybrid 1PL MLE		
		+	0	-
Separate 3PL EAP	+	147	36	0
	0	8	259	7
	-	0	40	143

Notes for Tables 4-6: School classifications are based upon estimated posterior means and SDs of school effects as specified in the layered model. The category "+" represents a school with an estimated value-added effect that remains above 0 after one posterior SD has been subtracted from its posterior mean. The category "0" represents a school with an estimated value-added effect that crosses 0 after one posterior SD has been subtracted from or added to its posterior mean. The category "-" represents a school with an estimated value-added effect that remains below 0 after one posterior SD has been added to its posterior mean.

[1]The layered model is the statistical machinery that underlies the Tennessee Value-Added Assessment System, which outside of the context of its usage in Tennessee is known more generally as the Educational Value-Added Assessment System. The layered model was developed by Dr. William Sanders, and is arguably the most established and well-known value-added model being used for the purposes of educational accountability.

[2] Other designs that would in principle also support an IRT-based approach include what Holland and Dorans (2006, pp. 197-201) describe as the single group, equivalent-groups, and counterbalanced designs. These designs are much less commonly enacted in operational testing programs because they require multiple administrations of the same test, which can be both costly and create problems maintaining test security.

[3] Our use of the term linking is consistent with the terminology established by Mislevy (1992), Linn (1993) and Holland & Dorans (2006). Linking is more general than the term equating, which refers to the linking of scores on alternate forms of an assessment that are built to common content and statistical specifications. Holland & Dorans (2006) distinguish between three forms of linking: predicting, aligning and equating. Under their taxonomy, vertical scaling is a form of score aligning in which the underlying tests have similar constructs and reliability but differing difficulty and test-taking populations.

[4] The EAP estimator is an example of a shrunken ability estimate.

[5] For philosophical debates over the meaning of measurement in the context of IRT models, see Wilson (2004), Thissen & Wainer (2002), and Wright (1997).

[6] For an exception, see Thissen & Orlando (2001) for an approach in which item parameters are estimated using the 3PLM, but proficiency estimates are based upon summed scores. In this case, there would be a one to one relationship to raw scores and scale scores when using a more complex model.

[7] 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.

[8] This assumption has been called into question in the context of estimating teacher effects. We have recently explored this issue in the context of school effects and found this to be a rather thorny issue (Briggs & Weeks, 2008). In short, while we found some evidence that school effects, like teacher effects, do not persist undiminished over time, these parameters are very difficult to identify and estimate when schools are the units of interest.

[9]The patterns of the results we present here were consistent across both longitudinal cohorts. We present the results for just the first cohort due to space constraints.

[10] Again, results for are presented for the first cohort only.

[11] These totals represent sums of off-diagonal values.