

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

Derek C. Briggs

Jonathan P. Weeks

Edward Wiley

University of Colorado, Boulder

June, 2008

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 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)?

Addressing these issues is clearly important, but they must be balanced by equal scrutiny of the dependent variables used on the left hand side of the equation. When value-added models are applied to estimate teacher or school effects across multiple grades over time, a central assumption is 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 are linked statistically to create what is known as a vertical scale. This linking process is known as “calibration”¹. The importance of the assumption that test scores have been vertically scaled is well understood by most psychometricians, but is easy to take for granted in applications of value-added modeling. For instance, Martineau (2006) demonstrated mathematically how violations of the vertical scale assumption of unidimensionality can lead to dramatic distortions of value-added estimates. The sensitivity of vertical scales to different linking designs and calibration approaches has also received greater attention in recent years (c.f., Tong & Kolen, 2007; Keller, Skorupski, Swaminathan, & Jodoin, 2004; Karkee, Lewis, Hoskens, Yao & Haug, 2003; Hanson & Beguin, 2002; Kim & Cohen, 1998). A key issue then, is whether longitudinal interpretations of student score changes are sensitive to the decisions made in creating a vertically linked scale.

The purpose of this paper is to conduct a large-scale empirical sensitivity analysis. 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?

¹Calibration is distinct from “equating”. According to Mislevy (1992) and Linn (1993), for two tests that measure the same construct, the term equating refers to the linking of scores on alternate forms of an assessment that are built to common content and statistical specifications, while the term calibration is used when scores are linked for tests with different levels of reliability or difficulty.

2. What impact do different vertical scaling approaches have on 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 calibration 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 model²” (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

In item response theory (IRT: Lord, 1980; Lord & Novick, 1968), an examinee’s score on a test item is modeled probabilistically as a function of the examinee’s latent ability 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 for what is known as the three-parameter

²The 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.

logistic model (3PLM: Birnbaum, 1968) can be written in the following form

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

where

θ_p = latent ability (measured in logits, which are the log of the odds of a correct item response),

α_i = item discrimination (slope at the location of item difficulty),

β_i = item difficulty (value of theta at the ICC inflection point),

γ_i = lower asymptote (guessing parameter),

D = a scaling constant (the value 1.7 is typically used. See Birnbaum, 1968 for more information).

The 3PLM is a general model for dichotomous items. Figure 1 is a graphic representation of the 3PLM, and illustrates how the probability of a correct item response is modeled for hypothetical examinees with abilities ranging from -4 to 4 logits for an item with a difficulty of $\beta = 1$, discrimination of $\alpha = 0.7$, and guessing parameter equal to $\gamma = 0.2$. For the two-parameter logistic model (2PLM: Birnbaum, 1968) the guessing parameter is constrained to zero, and for the one-parameter model (1PLM) the 2PLM is constrained so that the discrimination parameters for all items are equal. Though it has different historical and philosophical origins, the Rasch model (Rasch, 1960) is itself a special case of the 1PLM where all of the item discriminations are constrained to equal one.

Insert Figure 1 here

Constructed response items, generally referred to as polytomous items, can also be modeled using IRT. The category response curves for what is known as the generalized partial

credit model (GPCM: Muraki, 1992) take the following form

$$P(X_{pi} = v | \theta_p, \alpha_i, \beta_{iK_i}) = \frac{\exp\left[\sum_{v=1}^k D\alpha_i(\theta_p - \beta_{iv})\right]}{\sum_{h=1}^{K_i} \exp\left[\sum_{v=1}^h D\alpha_i(\theta_p - \beta_{iv})\right]} \quad (2)$$

where

θ_p = latent ability

α_i = item discrimination

β_{iK_i} = step difficulty for item i with K response categories

D = a scaling constant

Figure 2 is a graphic representation of the GPCM, and illustrates how the probability of a correct item category response is modeled for hypothetical examinees abilities ranging from -4 to 4 logits for an item with three response categories, a discrimination of 0.7, and two step difficulty parameters equal to 0.4 and 1. 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 on large-scale assessments include a mixture of dichotomous and polytomous items they are commonly modeled using a combination of the 3PLM and GPCM or the 1PLM and PCM.

The decision about which IRT model to use in establishing a vertical scale may be made

for statistical, pragmatic and/or even philosophical reasons³. 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 ability. On the other hand, the more parsimonious models 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, the rank orders of student raw scores and 1PLM/PCM scores are equivalent. The same is not true for the 3PLM and GPCM.

Insert Figure 2 here

One fundamental assumption all IRT models make is that of local independence, which posits that conditional on an examinee's latent ability (θ_p), the item responses within a given test should be statistically independent observations. All IRT models also share a common property when the specified model(s) fit the data: *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, for example, 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 it—even though in an absolute sense 5th grade students should generally have higher reading ability than 4th grade students. However, because of another property of IRT, *scale indeterminacy*, it is possible that the estimated item parameters will differ for the same

³ For philosophical debates over the meaning of measurement in the context of IRT models, see Wilson, 2004; Thissen & Wainer, 2002; and Wright, 1997.

item(s) administered on two different tests. In IRT models either item or examinee parameters must be fixed in order for the model(s) to be identified. In practice, the ability distribution at each grade is usually specified to be standard normal. Applying this constraint may give the appearance that the item parameters from two tests with common items are different, but this may just reflect relative differences examinee population distributions. These differences can be resolved using a linear transformation (i.e., transforming the item parameters for a given test so that the parameters for the common items administered on both tests are the same). On the basis of these properties 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 linking strategy is formally known as “common-item nonequivalent groups” linking (Kolen & Brennan, 2004).

In general, there are two IRT approaches used to create a vertical scale across two or more different grades: separate or concurrent calibration. Under separate calibration, parameters for items in adjacent grades are first estimated separately using the same IRT models. As an illustration, imagine two reading tests administered at the end of a school year, one test is given to grade 4 students, the other to grade 5 students. Each test has 50 multiple-choice items; 15 items are common to both grades, and 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 parameter invariance, the discrimination, difficulty, and guessing parameter of each common item ($i = 1, \dots, 15$) on the grade 5 test ($\alpha_{i5}, \beta_{i5}, \gamma_{i5}$) will have the following relationship with the corresponding parameters for the same items on the grade 4 test ($\alpha_{i4}, \beta_{i4}, \gamma_{i4}$)

$$\alpha_{i4} = \frac{\alpha_{i5}}{A}$$

$$\beta_{i4} = A\beta_{i5} + B$$

$$\gamma_{i4} = \gamma_{i5}$$

where A and B are the linking constants used for the linear transformation. Because the effect of scale indeterminacy is the same for all items on the grade 5 test (both common and unique), 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. In the context of creating a longitudinal vertical score scale, the ability estimates for examinees on the grade 5 scale can be placed onto the grade 4 scale using the following transformation $\theta_{p5}^* = A\theta_{p5} + B$ where θ_{p5}^* is the transformed score. Separate calibration can also be applied to more than two grades through “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 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 practice, linking constants are unknown. Various estimation methods have been proposed, but the Stocking- Lord method (Stocking & Lord, 1983), which minimizes the difference between test characteristic curves represented by the common items between grades, has become the de facto standard.

Under concurrent calibration, all item parameters for all grades are estimated in one step during which different underlying population distributions are specified for the nonequivalent groups of students taking the tests across grades (Bock & Zimowski, 1997). 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 common items providing the link to anchor the two tests.

Separate and concurrent calibration approaches each have strengths and weaknesses. Separate calibration is easy to implement, but because it is unlikely that linking constants are estimated without error, additional error is usually introduced into the transformed scale. This is particularly the case as the transformed grade departs further from the base grade (Kim, Lee, & Kim, 2008). On the other hand, with the concurrent approach only one calibration run is needed to estimate all the parameters and create the vertical scale. In this regard there is no real linking error. Nevertheless, separate calibration is not without advantages of its own: 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, when the definition of the construct measured changes across grades, concurrent estimation can introduce bias throughout the entire scale, whereas separate calibration may mitigate such bias by relying only on pairwise linking (Béguin & Hanson, 2001; Béguin, Hanson, & Glas, 2000).

For the present study, we developed scales using a separate approach, and a modification of a purely concurrent approach. In most applications of IRT calibration, a cross-section of examinees in a single year are tested across a vertical grade span; this becomes the basis for establishing a vertical link across the tests. In contrast, our data relies on two longitudinal cohorts of students—the same students are included in the dataset on multiple occasions. This creates possible violations of the IRT local independence assumption. For this reason, instead of using a purely concurrent approach, we used a hybrid method where item parameters for the within-grade/across-year assessments were estimated concurrently and then separate calibration

was used to create the across-grade scales. We illustrate the differences between these approaches graphically in the next section.

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) estimation. 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 as Bock and Mislevy (1982, p. 439) note “the EAP estimator has minimum mean square error over the population of ability and, in terms of average accuracy, cannot be improved upon.”

Methods

Data

We obtained longitudinal item responses from the Colorado Department of Education for two cohorts of students on the Colorado Students Assessment Program (CSAP) reading test. The structure of this data is shown in Figure 3 below.

Insert Figure 3 here

We obtained the data for two longitudinal cohorts because the vertical linking design employed by the state’s test contractor, CTB/McGraw-Hill, includes no common items between the tests given to students in the same cohort in adjacent years. As a result, we could only create a vertical scale by first linking tests for adjacent grades in the same year, and then linking tests for the same grade in adjacent years. An unexpected complication was the fact that CTB does not always include common items across adjacent grades in the same year, or across the same grade in adjacent years. Given this, we were fortunate to discover that there were in fact common items in adjacent grades and years for our two student cohorts taking the reading tests in grades 3 through 7 from 2003 to 2006. The linking design for the CSAP reading test 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. A limitation of this design is that for some of the links—particularly between the years 2005 and 2006 for grade 6, the number of common items available (9) is relatively small.

Insert Table 1 here

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 (this number includes charter schools, but excludes private 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.

Vertical Scaling

We created eight vertical scales that differ with regard to three variables: (1) the IRT model used to estimate item parameters, (2) the calibration 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, which ranges from third to tenth grade, is currently based on the combination of variables represented in cell number 1: 3PLM/GPCM, separate calibration, and EAP score estimation. However, the scales represented by cells 2-7 would be psychometrically defensible alternatives. The vertically scaled scores from each of these combinations are subsequently used as different longitudinal outcome variables for the layered value-added model.

Insert Table 2 here

The two different calibration approaches taken in this study are illustrated graphically in Figures 4 and 5. In Figure 4, each oval represents a separate calibration 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 5, separate calibrations are performed for tests across grades in the same year (ovals), but prior to the separate calibrations a concurrent calibration is performed for tests across two years in the same grade (rectangles). Under the separate approach, item parameters for each grade were first estimated independently. Next, using the Stocking-Lord method (Stocking & Lord, 1983), linking constants were estimated for the various within-year/across-grade and within-grade/across-year pairs illustrated in Figure 4. These constants were used to transform both item parameters and estimates of latent proficiency, θ , onto the grade 3 scale using chain linking. The estimation of linking constants and the chain linking transformation were accomplished using the R package plink (Weeks, 2007). Under the hybrid approach, item parameters were first estimated for each of the within-grade/across-year assessments concurrently. Next, the Stocking-Lord method was used to estimate linking constants for the across-grade linkages shown in Figure 5. The item parameters and ability estimates for these scales were then placed onto the grade 3 scale using chain linking. All of the item parameters were estimated using the IRT Command Language program (ICL; Hanson, 2002).

Insert Figures 4 and 5 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 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:

$$Y_{it} = \mu_t + \sum_{t^* \leq t} \alpha_{tt^*} \boldsymbol{\theta}_{t^*} + \varepsilon_{it}. \quad (3)$$

In equation 3, Y_{it} represents the CSAP reading test score for student i in year t , $t = 1, \dots, T$, and the parameter μ_t denotes the grand test score mean for a given year. The vector $\boldsymbol{\theta}_{t^*}$ represents the collection of school effects⁴ for each year, and the parameter α_{tt^*} captures the persistence of the school effects $\boldsymbol{\theta}_{t^*}$ in year t (given that $t^* \leq t$). Finally, ε_{it} represents the test score residual associated with student i in year t . Both $\boldsymbol{\theta}_{t^*}$ and ε_{it} are assumed to be independent latent random variables, where $\varepsilon_{it} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\theta}_{t^*} \sim N(\mathbf{0}, \boldsymbol{\tau})$. 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 $t^* \leq t$)⁵.

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

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

⁴ 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.

⁵ 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, we found evidence that school effects, like teacher effects, do not persist undiminished over time when longitudinal data is being analyzed.

We note the following about the school-level parameters in these equations. First, the parameter vectors $\{\theta_{04}, \theta_{05}, \theta_{06}\}$ 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 θ_{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, we note that 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 6 and 7 for the first of our two longitudinal student cohorts⁶: 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 calibrate the vertical scales (separate vs. hybrid). All scales that were created using separate calibration exclusively are denoted graphically by solid lines, and the scales created using the hybrid calibration are denoted with dotted lines.
3. The difference between approaches used to estimate student-level scale scores (EAP vs. MLE). All scales that were created with EAP estimation are denoted graphically with asterisk markers, and the scales created with ML estimation are denoted with square markers.

Insert Figures 6 and 7 here

⁶ 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.

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 6. For most of the vertical scales, the growth in scale score means from grade to grade appears somewhat nonlinear, decelerating over time. As would be expected, the choice of estimation approach (EAP vs. ML) has no significant impact on mean estimates once the underlying IRT model and calibration approach are held constant. On the other hand, the magnitude of growth does differ substantially as a function of the underlying IRT model and calibration approach. Scales created using the 3PLM/GPCM and separate calibration combination give the impression of the most growth; scales created using the 1PLM/PCM and either separate or hybrid calibration give the impression of the least growth. In effect, 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. Interestingly, however, when 3PLM/GPCM estimates are calibrated using the hybrid approach rather than separate approach, the observed growth trajectory is shifted downward.

Much of the differences in growth by scale are artifacts of each scale's variability, as can be seen in Figure 7. 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 use of separate calibration increases variability relative to hybrid calibration. While the choice of IRT model has the largest impact on scale variability, the choice of calibration and estimation approach still has a substantial impact, especially when combined with the 3PLM/GPCM. One interesting finding concerns the changes in variability across grades.

While the variability across all scales decreases or remains roughly constant after the base year grade, variability increases slightly across the subsequent grades in the longitudinal span for the 3PLM/GPCM scales, while for the 1PLM/PCM, the opposite is true—variability decreases slightly.

One upshot of these findings 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 8, which contrasts two extremes: a vertical scale based on separate calibration, the 3PLM/GPCM and ML estimation (“s3m”) and a vertical scale based on hybrid calibration, the 1PLM/PCM and EAP estimation (“h1e”). In terms of the three decisions used to establish each vertical scale in each case, the s3m scale maximizes score variability while the h1e scale minimizes it.

Insert Figure 8 here

A good way to evaluate year to year growth along a given scale while adjusting for the variability of the underlying scale is to standardize the mean differences by computing an effect size. According to Yen (1986) this statistic can be defined as

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

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_{upper}^2 and σ_{lower}^2 represent the respective variance for the scores in

each grade or year. The effect sizes that correspond to the growth from grades 3 to 4, 4 to 5 and 5 to 6 are shown in Figure 9. Ideally, one would hope to find no substantive differences in year to year effect sizes as a function of the scale, but this is not the case here. For each longitudinal grade comparison, effect sizes can differ by as much as 10 to 20 percent of the pooled standard deviation across grades. While there is a consistent pattern of effect size decreases across scales over the first three years, the effect sizes for scales created under the 1PLM/PCM and hybrid calibration increase. In contrast, the effect sizes continue to decrease for the corresponding scales created under separate calibration. These differences between the separate and hybrid scales can probably be explained by the small number of common items available to link scores from grade 6 in 2005 to scores from grade 6 in 2006 (linking the two grade 6 tests was a necessary precursor to placing the 2005 grade 5 scores and 2006 grade 6 scores on the same scale). Under separate calibration, linking constants are estimated solely as a function of the available common items. If these common items are not representative of the underlying construct being measured, the resulting estimates of linking constants may be biased. Conversely, in our hybrid calibration approach the grade 6 link was established using concurrent calibration. In the latter approach, both unique and common items are used to establish an implicit link between tests, hence there appears to be less sensitivity to the small number of available common items.

Insert Figure 9 here

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”⁷. 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 two posterior standard deviations above zero and “below average” if its estimated effect is more than two standard deviations below zero.

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..

⁷ Again, results for are presented for the first cohort only.

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, selection of calibration method makes little difference in the percent of schools classified as above or below average. In grades 4 and 6, scales based upon the combination of 3PLM/GPCM and EAP with a separate calibration approach classify the most schools. In grade 5, scales based upon the combination of 3PLM/GPCM and EAP with a hybrid calibration approach classify the most schools. 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 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 as 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 other⁸. If sanctions or rewards are attached to these classifications, the choice of scale can clearly have important ramifications.

⁸ These totals represent sums of off-diagonal values.

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 growth 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. The recent proliferation of value-added modeling approaches provides an appealing alternative, but often at the cost of great statistical complexity and misguided causal inferences (Braun, 2005; Briggs & Wiley, 2008; Raudenbush, 2004; Rubin, Stuart & Zanatto, 2004)

One key assumption that has been often 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 calibration method 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 is most likely to be problematic.

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. On the basis of some unexpected differences by calibration approach found for the link with the fewest common items (grade 6 in 2005 and 2006), it appears that the number of common items can have differential impact on the resulting scale as a function of whether a separate or concurrent calibration approach is used. The theoretical explanation for this finding is unclear at this point, in part because there are no examples in the literature in which a vertical scale has been established using longitudinal data. As far as we know, all vertical score scales established and maintained by testing companies rely upon cross-sectional data to perform calibrations across multiple grades within a single year. The ideal check on the methods we have used in this study would involve the full panel data for grades 3 through 7 between 2003 and 2006, but this data was not available to us. Finally, our analysis has only considered differences among vertical scales in the subject of reading. However, while there might be some small differences in the vertical scales created for the

subjects of math and science, we would not expect to see significant changes in the scale to scale patterns with respect to subsequent value-added estimates.

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 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. 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.

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

- 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.
- 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.
- 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., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 34, 197-211.
- Bock, R. D., & Zimowski, M. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433-448). 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.

- Hanson, B.A. (2002) IRT command language. Monterey, CA. (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.
- 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.
- Keller, L., Skorupski, W., Swaminathan, H., and Jodoin, M. (2004) *An evaluation of item response theory equating procedures for capturing changes in examinee distributions with mixed-format tests*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, April 2004, San Diego, CA.
- 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.
- 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.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

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

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

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

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., 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.

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.

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.

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.

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

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.

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>

Wilson, M. (2004). On choosing a model for measuring. In *Introduction to Rasch Measurement*, E. V. Smith and R. M. Smith (eds.), JAM Press.

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

Yen, W. (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

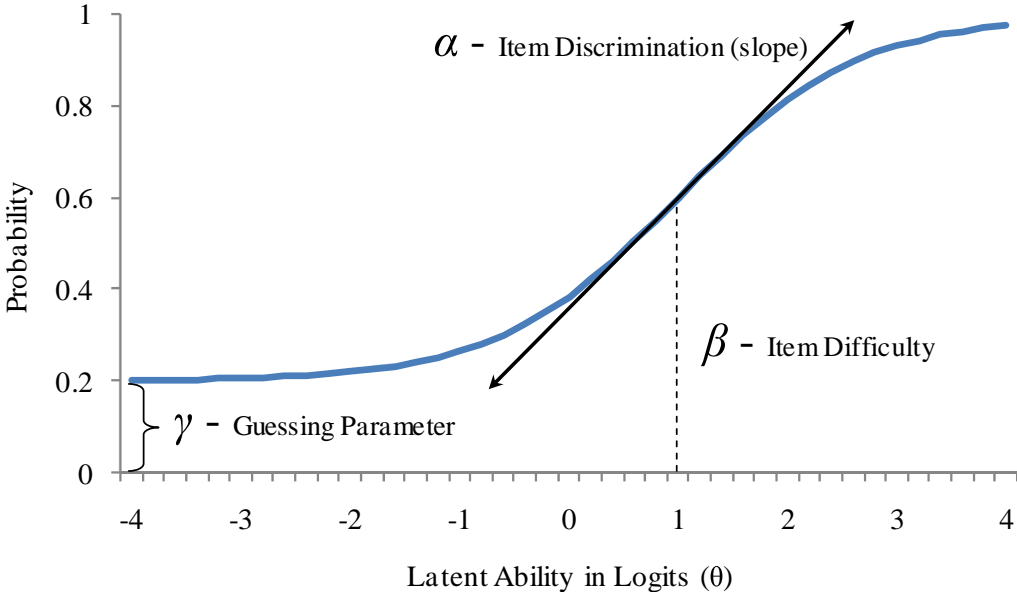


Figure 2. GPCM Category Characteristic Curves

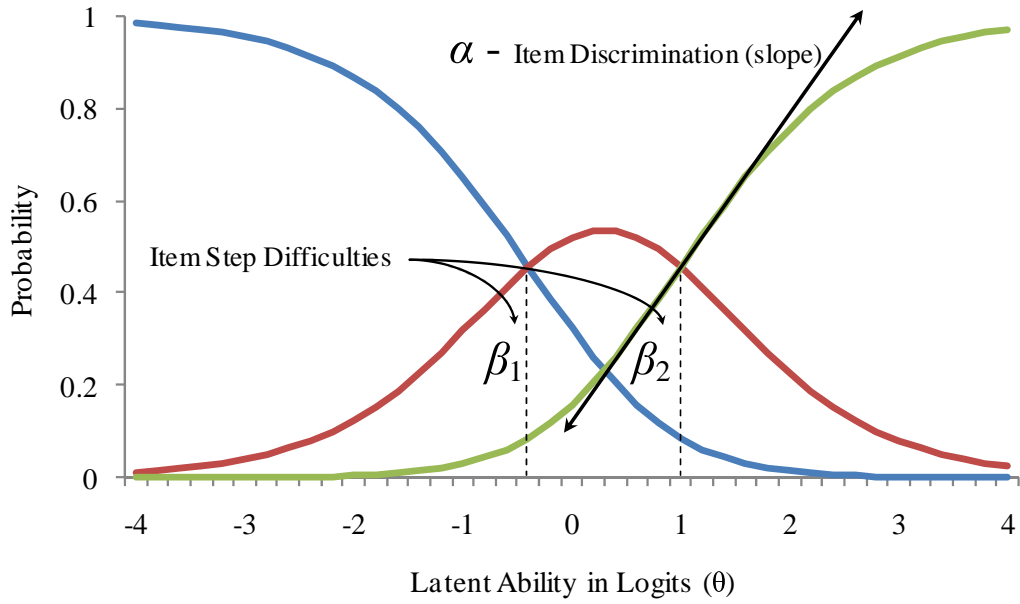


Figure 3. Cohort Files Obtained from CDE for all Students in State of Colorado

Grade Cohorts	Year			
	2003	2004	2005	2006
Grade 3 Reading	3			
Grade 4 Reading	4	4		
Grade 5 Reading		5	5	
Grade 6 Reading			6	6
Grade 7 Reading				7

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 4. Separate Calibration Approach

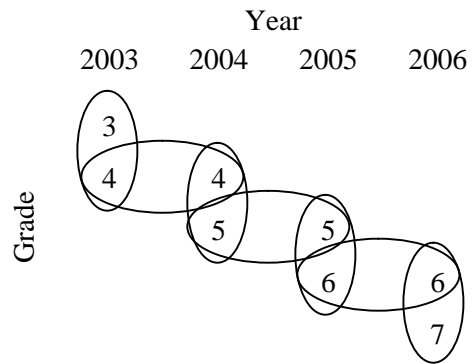


Figure 5. Hybrid Calibration Approach

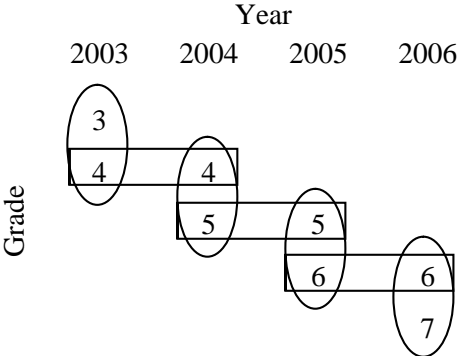


Figure 6. Growth along CSAP Reading Score Scale from 2003 to 2006

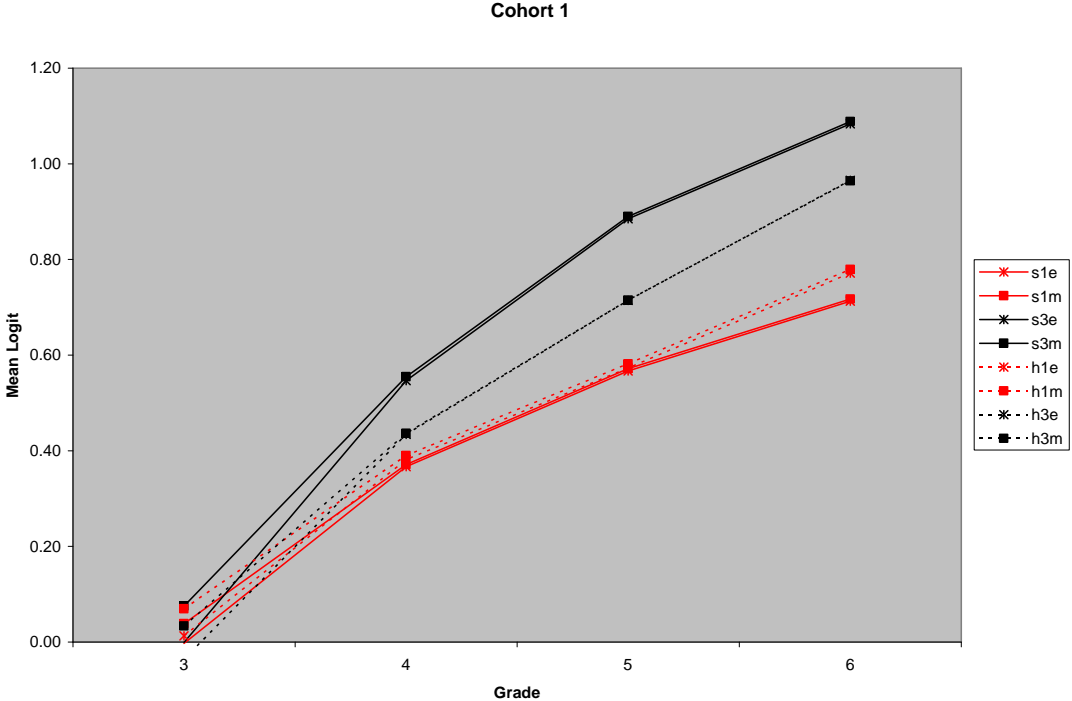


Figure 7. Variability of CSAP Reading Score Scale from 2003 to 2006

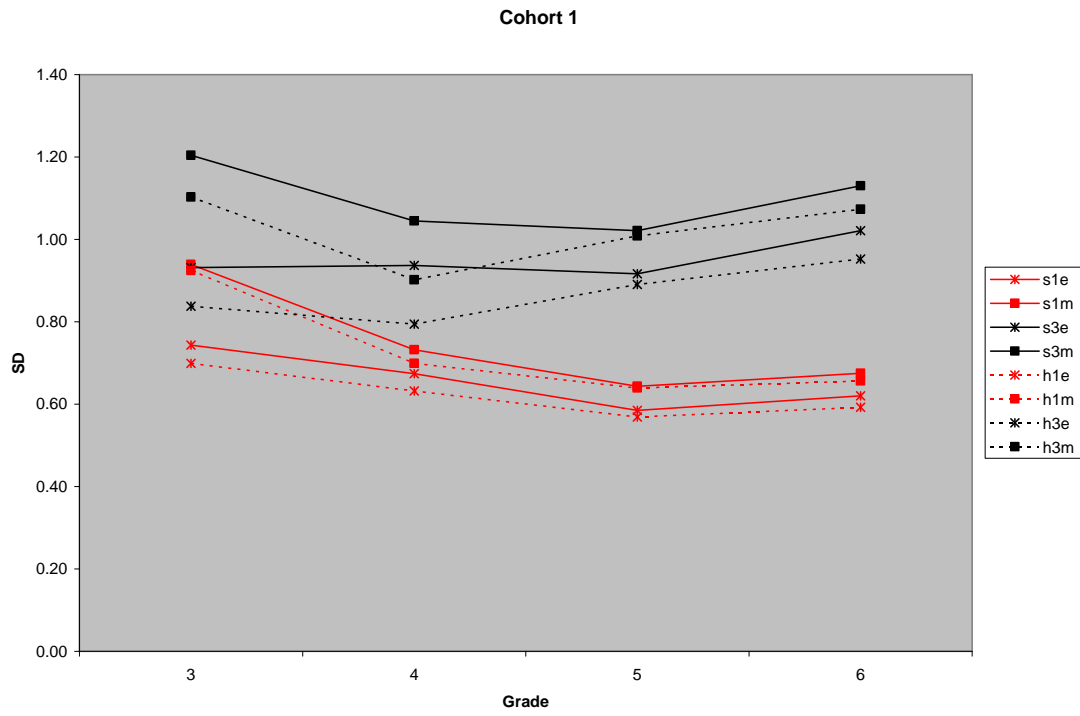


Figure 8. Comparing Extremes in Vertical Scales

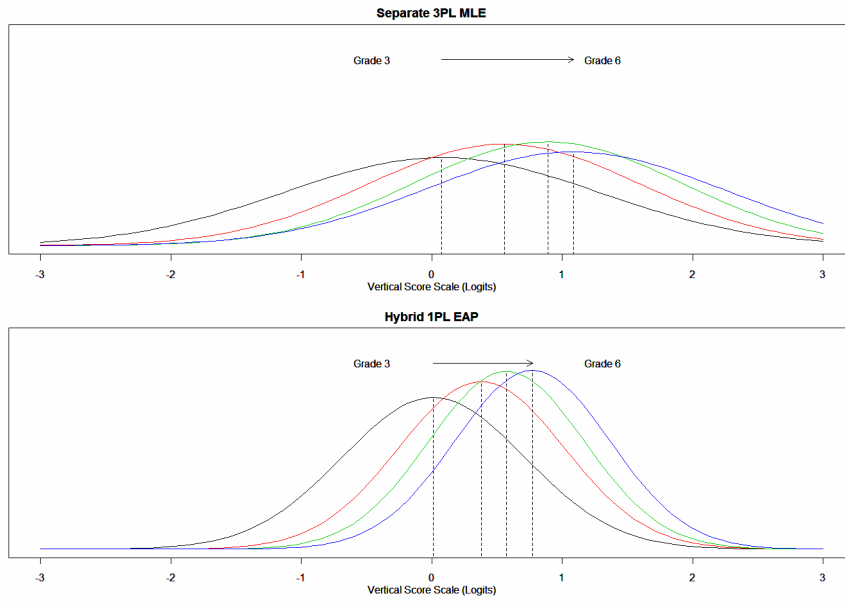


Figure 9. Effect Sizes of Growth by Scale for Cohort 1

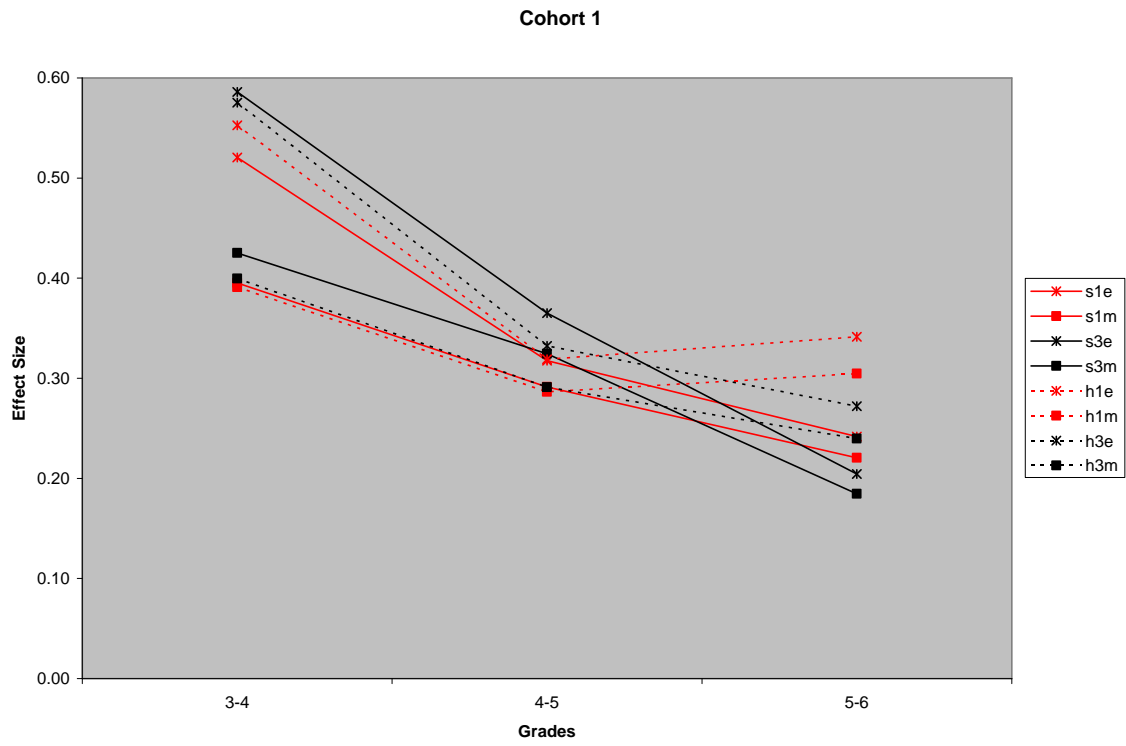


Table 3: Comparison of School Classifications by Underlying Vertical Scale

		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 “+” represents a school with an estimated value-added effect that remains above 0 after two posterior SDs have been subtracted from its posterior mean. The category “0” represents a school with an estimated value-added effect that crosses 0 after two posterior SDs have 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 two posterior SDs have 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 two posterior SDs have been subtracted from its posterior mean. The category “0” represents a school with an estimated value-added effect that crosses 0 after two posterior SDs have 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 two posterior SDs have been added to its posterior mean.

This document was created with Win2PDF available at <http://www.win2pdf.com>.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.