The Impact of Vertical Scaling Decisions on Growth Projections

Derek C. Briggs

Jonathan P. Weeks

Edward Wiley

University of Colorado, Boulder

March, 2008

Paper presented at the annual meeting of the National Council for Measurement in Education Annual Conference, March 26, 2008, New York, NY.

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

Introduction

In educational research, a constant methodological challenge is finding ways properly evaluate changes in student achievement, and to distinguish the influence of teacher and schools on such changes. Two related statistical models have garnered increasing attention for these purposes: Growth models and value-added models. The key feature of all growth models is the availability of longitudinal data such that changes in student achievement can be parameterized over time. Given the availability of teacher or school-level variables, most growth models can also be conceptualized as value-added models. However, the converse is not necessarily true. In most value-added models, growth is not formally parameterized, and the only purpose served by student achievement data is to estimate residualized contributions attributable to teachers and/or schools.

The November 2005 announcement of the Growth Model Pilot Program (GMPP) (Spellings, 2005) opened the door to fundamental changes to the No Child Left Behind Act of 2001 (NCLB). Under the GMPP, ten states are being selected to implement proposed growth models on a trial basis, allowing individual student growth trajectories to factor into accountability calculations for the first time (U.S. Department of Education, 2005). The incorporation of student growth into accountability—while not opening the door to the use of value-added modeling techniques as part of NCLB—places increasing emphasis on the change in student achievement over the status of that achievement at any single point in time (Carlson, 2006; Hill, et al., 2005, Goldschmidt, et al., 2005).

A central assumption underlying growth models is that test scores have been vertically scaled such that they have a consistent interpretation over time. To create a vertical scale, scores from two or more tests are linked statistically so that scores from the tests can be expressed on a common scale. This linking process is known as calibration[1]. The importance of the assumption that test scores have a consistent vertical scale is well understood by most psychometricians, but seems to be taken for granted by statisticians in many applications of growth models and their value-added extensions. For example, Martineau (2006) has demonstrated mathematically how violations of the vertical scale assumption of unidimensionality can lead to dramatic distortions in value-added estimates. In addition, the sensitivity of vertical scales to different linking designs and calibration approaches has also begun to receive greater attention in recent years (c.f., Tong & Kolen, 2005; Keller, Skorupski, Swaminathan, & Jodoin, 2004; Karkee, Lewis, Hoskens, Yao & Haug, 2003; Hanson & Beguin, 2002; Kim & Cohen, 1998). A key issue then, is whether vertical scaling is stable enough as a measurement enterprise to support consistent and reasonably precise longitudinal interpretations when subsequently applied in the context of a growth model.

The purpose of this paper is to evaluate the sensitivity of growth modeling results to the way an underlying vertical scale has been established. We accomplish this by analyzing longitudinal item-level data with both student and school-level identifiers over time in the state of Colorado. We use this data to address two principal research questions:

[1] Calibration is distinct from equating. As developed by Mislevy (1992) and Linn (1993), the term equating refers to linking 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 on test that are intended to measure the same construct but with different levels of reliability or difficulty (Kolen, 2004).

1. What is the sensitivity of a longitudinal score scale to the way the test scores have been vertically scaled?
2. What impact do different IRT-based vertical scaling approaches have on projections of growth in student achievement?

The basic strategy taken here is to create different vertical scales on the basis of choices made for three key variables: IRT modeling approach, calibration approach and student proficiency estimation approach. Combinations of among these three variables leads to eight different vertical scales. Each scale represents a methodological approach that is in some sense defensible. Of interest at this stage are potential differences in means and standard deviations among the different vertical scales from year to year. We next use the longitudinal values of each scale as the outcome variable in a relatively simple value-added growth model. Of interest at this stage are comparisons among the different fixed effect estimates of growth, and empirical Bayes estimates of student and school-level growth parameters. Our findings suggest that growth projections may in fact be quite sensitive to choices made in the development of a vertical scale.

Methods

Data

We obtained longitudinal item responses from the Colorado Department of Education for four cohorts of students on the Colorado Students Assessment Program (CSAP) tests of math and reading. The structure of this data is shown in Figure 1 below.

Figure 1. Cohort Files Obtained from CDE for all Students in State of Colorado

			Year		
Grade Cohorts	2003	2004	2005	2006	2007
Grade 3 Reading	3				
Grade 4 Reading	4	4			
Grade 5 Reading		5	5		
Grade 6 Reading			6	6	
Grade 7 Reading				7	7
Grade 8 Reading					8

			Year		
Grade Cohorts	2003	2004	2005	2006	2007
Grade 5 Math	5				
Grade 6 Math	6	6			
Grade 7 Math		7	7		
Grade 8 Math			8	8	
Grade 9 Math				9	9
Grade 10 Math					10

It was necessary to obtain two longitudinal cohorts for each test subject because the vertical linking design employed by CTB includes no common items between the tests given to students in the same cohort in adjacent years. As a result, we can 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 additional and 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. By luck, there were common items in adjacent grades and years for our two student cohorts taking the reading tests in grades 3 through 8 from 2003 to 2007. Unfortunately, this was not the case for the two student cohorts taking the math tests in grades 5 through 10 from 2003 to 2007. As a result, at this point (until we can get data for another longitudinal cohort of students who took the CSAP math test in different grades from 2003 to 2007) we can only report the results from the vertical scaling of the CSAP reading test.

The CSAP reading tests used to establish different vertical scales contained a mix of multiple-choice (MC) and constrained-response (CR) items. In grade 3 the test consisted of 41 MC items and 7 CR; 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 9 to 20 and 0 to 4. The linking design for the CSAP reading test is summarized in Table 1 below.

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

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

For each of the nine grade by year combinations used to establish a vertical scale in what follows (from grades 3 to 7), there were on average roughly 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

The testing structure presented in Table 1 represents what is known as a common item non-equivalent group linking design (c.f., Kolen & Brennan, 2004). The development of a vertical score scale from this design requires a psychometric model for placing test scores onto a common scale, and a calibration and proficiency estimation approach to be used in conjunction with the psychometric model. We describe the different choices we have made with respect to these three variables below. Table 2 provides an overview of the eight different vertical scales that were created in this study, represented by the cells 1-8.

Table 2. IRT-Based Vertical Scaling Models

Item Response Model		Linking Approach	
		Separate Calibration	Hybrid Calibration
EAP Scale Scores	Different item weights[1]	1	2
	Equal item weights[ii]	3	4
MLE Scale Scores	Different item weights[1]	5	6
	Equal item weights[ii]	7	8

[1] Represented by application and extension of the Three Parameter Logistic Model (3PLM) and Generalized Partial Credit Model (GPCM).

[ii] Represented by application and extension of the One Parameter Logistic Model (1PLM) and Partial Credit Model (PCM).

Psychometric Models

The CSAP reading tests items are scaled by the state of Colorado's test contractor CTB using a combination of the three parameter logistic model (3PLM) for dichotomous items, and the generalized partial credit model for polytomous items (GPCM). However, it would also be conceivable to scale these items using a combination of the one parameter logistic model (1PLM) for dichotomous items, and the partial credit model for polytomous items (PCM). One important distinction between these two different IRT model combinations comes in the role played by test items in the subsequent estimation of student scale score. In the 3PLM/GPCM combination items are weighted by their discriminating power; in the 1PLM/PCM combination they are unweighted because items are constrained to discriminate equally.

Calibration Approaches

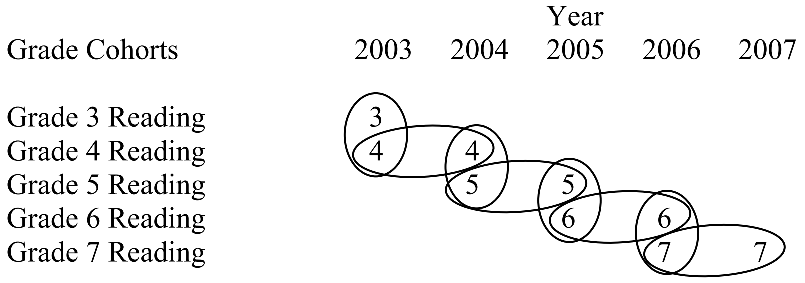
Two of the most widely used approaches to create a vertical scale across two or more different tests involve separate or concurrent calibration. Under separate calibration, parameters for items in adjacent grades are estimated separately, and then a single vertical scale is established in a subsequent step using the common items between grades. That is, given two different proficiency scales corresponding to, for example, grades 5 and 6, scores for student n $\{n = 1, \ldots, N\}$ in grade 6 can be placed onto the scale of scores for grade 5 using the linear transformation $\theta_n^* = A\theta_n + B$. Estimates for the linking parameters A and B can be obtained using and the Stocking and Lord (1983) algorithm, which minimizes the difference in test characteristic curves represented by the common items between grades. Another defensible vertical scaling approach would be to use concurrent calibration. In concurrent calibration, all item parameters for all grades are estimated in one step with a multigroup IRT model and placed on a single vertical scale. With respect to summary measures of grade to grade growth and variability in 2002

CSAP math and reading test scores, Karkee et. al. (2003) found no substantial difference between the separate and concurrent approaches in the aggregate.

In this study we are comparing a separate calibration approach to an approach that is essentially a hybrid of the separate and concurrent approaches. Under the separate approach, item parameters for each grade were first estimated independently, and then placed onto the grade 3 scale in a chained manner using the Stocking and Lord algorithm to estimate the appropriate linear transformation parameters from grade to grade. These parameters were also used to transform estimates of latent proficiency, θ, onto the same vertical scale from grades 3 through 8. For the separate calibration, item parameters were estimated for each grade separately using IRT Command Language (ICL; Hanson, 2002), and then linked together using the R package plink (Weeks, 2007).

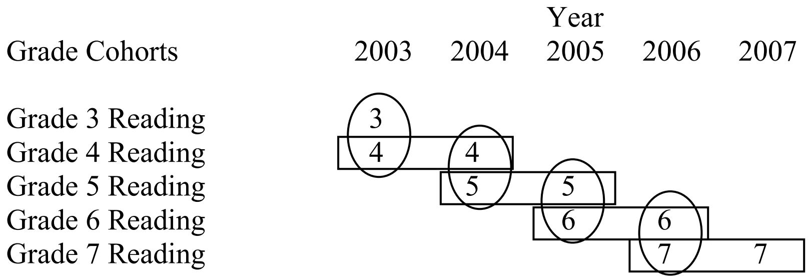
The procedure through which separate calibration establishes a vertical scale is illustrated in Figure 2, where each oval represents a separate calibration of tests across grades in the same year (vertical direction), or across years in the same grade (horizontal direction). Under the hybrid approach illustrated in Figure 3, separate calibrations are performed for tests across grades in the same year (ovals), but in between each separate calibration a concurrent calibration is performed for tests across two years in the same grade (rectangles).

Figure 2. Separate Calibration Approach



			Year		
Grade Cohorts	2003	2004	2005	2006	2007
Grade 3 Reading	3				
Grade 4 Reading	4	4			
Grade 5 Reading		5	5		
Grade 6 Reading			6	6	
Grade 7 Reading				7	7

Figure 3. Hybrid Calibration Approach



			Year		
Grade Cohorts	2003	2004	2005	2006	2007
Grade 3 Reading	3				
Grade 4 Reading	4	4			
Grade 5 Reading		5	5		
Grade 6 Reading			6	6	
Grade 7 Reading				7	7

Estimating Student Scale Scores

An important decision in either the separate or hybrid calibration approaches is the choice of method for the estimation of student scale scores after (or concurrent with) the estimation of item parameters. Two typical choices are maximum likelihood (ML) and expected a posteriori (EAP) estimates. There is a well-known bias-efficiency tradeoff between ML and EAP estimates of students scale scores. Because EAP estimates result from the weighted combination for each respondent of an empirical likelihood function and a prior distribution, they will be shrunken toward the population mean relative to ML estimates and thereby minimize measurement error. This implies greater variability in vertical scales comprised of ML estimates relative to EAP estimates. On the other hand, EAP estimates are biased; ML estimates are asymptotically consistent.

Growth Modeling

Growth was projected for students in the longitudinal cohort in grade 3 as of 2003 and in grade 7 as of 2007 were made on the basis of their performance on CSAP reading tests in grades 3-5 during elementary school. This reflects the type of "growth to standard" NCLB projection model that has been proposed (though not approved) by both the states of Hawaii and Oregon. Let , Y_{ijt}^{s} represent the test score *Y* for student *i* in school *j* at time *t* ($t = \{0,1,2\}$) on vertical scale *s*. We then specified the following mixed-effects model, also known as a hierarchical linear model.

$$Y_{ijt}^{s} = \beta_{00} + \zeta_{00i} + \theta_{00j} + (\beta_{01} + \zeta_{01i} + \theta_{01j})Time_t + \varepsilon_{ijt}$$

where

β_{00} = fixed effect intercept (grand mean),
β_{01} = fixed effect slope (average growth trajectory across all students and schools),
ζ_{00i} = random effect on the intercept for student *i*,
ζ_{01i} = random effect on the slope for student *i*,
θ_{00j} = random effect on the intercept for school *j*,
θ_{01j} = random effect on the slope for school *j*, and
ε_{ijt} = residual for student *i* in school *j* at time *t*.

The random effect parameters are assumed to take on the following distributions:

$$\varepsilon_{ijt} \sim N(0, \sigma^2)$$

$$\begin{bmatrix} \zeta_{00} \\ \zeta_{01} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} \right)$$

$$\begin{bmatrix} \theta_{00} \\ \theta_{01} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \psi_{00} & \psi_{01} \\ \psi_{10} & \psi_{11} \end{bmatrix} \right).$$

Note that an implicit assumption typical of such models is that each random effect parameter is independent across levels. We specified the model above for $s = \{1, 2, \dots, 8\}$, and obtained parameter estimates using the software HLM 6.05 (Raudenbush, Bryk & Congdon, 2008). In estimating these three level hierarchical models, we restricted the sample to only those students who were present in the same school in grades 3, 4 or 5 for two out three years between 2003 and 2005. This reduced the available student sample for our analysis from 65,599 to 40,690. The students that were excluded represent those that either switched schools, moved to a private school, or left the state. The sample of included students attended a total of 992 unique schools.

Results

Comparing Vertical Scales

There are two important statistics for summarizing a vertical score scale for any given grade and year: the mean and standard deviation (SD) of the scale. A third statistic, the effect size, is computed as a function of the mean and SD and is useful for comparing differences in the scale for any two adjacent years of grades. In the vertical scaling context, an effect size (Yen, 1986), is defined as

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

where $\overline{\theta}_{upper}$ and $\overline{\theta}_{lower}$ represent the mean scale score for the higher and lower grades or years in the scale, and σ^2_{upper} and σ^2_{lower} represent the respective SDs for the scores in each grade or year.

Table 3: CSAP Reading Vertical Scale Descriptive Statistics

			Separate Calibration				Concurrent Calibration			
			1PLM		3PLM		1PLM		3PLM	
	Grade	Year	EAP	MLE	EAP	MLE	EAP	MLE	EAP	MLE
Mean	3	2003	-0.003	0.038	-0.001	0.075	0.013	0.069	-0.035	0.034
	4	2003	0.314	0.320	0.473	0.483	0.323	0.333	0.332	0.338
	4	2004	0.367	0.371	0.547	0.555	0.381	0.390	0.434	0.436
	5	2004	0.604	0.607	0.927	0.926	0.610	0.618	0.741	0.739
	5	2005	0.567	0.572	0.885	0.890	0.573	0.582	0.714	0.715
	6	2005	0.718	0.720	1.088	1.078	0.712	0.717	0.878	0.862
	6	2006	0.713	0.717	1.083	1.089	0.771	0.779	0.965	0.964
	7	2006	0.769	0.773	1.161	1.156	0.848	0.855	1.103	1.096
SD	3	2003	0.743	0.939	0.931	1.204	0.699	0.925	0.838	1.103
	4	2003	0.698	0.767	0.933	1.038	0.659	0.738	0.773	0.887
	4	2004	0.674	0.732	0.937	1.045	0.632	0.699	0.794	0.902
	5	2004	0.598	0.651	0.873	0.988	0.581	0.647	0.762	0.871
	5	2005	0.585	0.644	0.916	1.021	0.569	0.639	0.890	1.009
	6	2005	0.642	0.700	0.984	1.109	0.630	0.700	0.941	1.084
	6	2006	0.620	0.675	1.021	1.130	0.593	0.656	0.953	1.073
	7	2006	0.590	0.641	1.004	1.134	0.578	0.640	0.970	1.100
Effect Size	3-4	2003	0.439	0.329	0.508	0.363	0.456	0.315	0.456	0.305
	4-4	2003-04	0.077	0.068	0.079	0.069	0.091	0.079	0.130	0.109
	4-5	2004	0.372	0.341	0.420	0.365	0.377	0.338	0.395	0.342
	5-5	2004-05	-0.062	-0.054	-0.047	-0.036	-0.065	-0.056	-0.032	-0.026
	5-6	2005	0.245	0.220	0.213	0.176	0.232	0.201	0.178	0.141
	6-6	2005-06	-0.008	-0.004	-0.005	0.010	0.097	0.092	0.093	0.095
	6-7	2006	0.092	0.084	0.076	0.059	0.130	0.117	0.143	0.121

Note: Means and SDs are expressed in logit units.

The means, SDs and effect size estimates for each of the eight vertical scales we created are summarized by grade and year combination in Table 3. For each statistic, there are three comparisons of interest (where each comparison is made while holding other two constant): (1) The difference between IRT models applied to estimate item parameters (1PLM/PCM vs. 3PLM/GPCM) (2) The difference between approaches used to calibrate the vertical scale (separate vs. hybrid).. (3) The difference between approaches used to estimate student-level scale scores (EAP vs. MLE). On the basis of the results above, we reach the following three general conclusions:

1. In an absolute sense, average growth as represented by differences in means appears larger when the 3PLM/GPCM is used to estimate item parameters instead of the 1PLM/PCM. This difference is accompanied by more variability in scale scores for the 3PLM/GPCM combination. However, when differences in means are standardized as effect sizes, differences along the scale as a function of IRT model, while still present, are less dramatic.
2. The means and SDs from separate and hybrid calibrations are generally quite similar when the underlying IRT model is the 1PLM/PCM. When the underlying model is the 3PLM/GPCM, the means and SDs under the hybrid approach are consistently smaller than those under the separate approach. When compared on the effect size metric, the two approaches never differ by more than 1/10 of an SD.
3. As one would expect, there is no difference in the means as a function of scale score estimation method, but the variability of scale scores estimated using EAPs is considerably smaller than the variability estimated using MLEs. As a consequence, effect sizes for scales based on EAP estimation are consistently bigger than effect sizes based on MLE estimation.

Comparing Growth Trajectories

The key results from applying our three-level hierarchical model to the scores from each vertical scale are presented in Table 4.

Table 4. Results from Application of Growth Model to Vertical Scales

	Separate Calibration				Hybrid Calibration			
	1PLM/PCM		3PLM/GPCM		1PLM/PCM		3PLM/GPCM	
	EAP	ML	EAP	ML	EAP	ML	EAP	ML
Fixed Effects								
β_{00}	-0.0144	0.0147	-0.0194	0.0343	0.0053	0.0471**	-0.0504***	-0.0033
β_{01}	0.2940***	0.2804***	0.4524***	0.4241***	0.2876***	0.2688***	0.3807***	0.3532***
Random Effects								
σ^2	0.0708	0.1079	0.1238	0.1923	0.0615	0.1023	0.09183	0.1618
τ_{00}	0.3866	0.6051	0.6093	1.0083	0.3388	0.5771	0.47485	0.8152
τ_{11}	0.0091	0.0416	0.0091	0.0504	0.0077	0.0419	0.01539	0.0513
τ_{01}	-0.0499	-0.1325	-0.0159	-0.1355	-0.0390	-0.1272	0.00313	-0.0933
ψ_{00}	0.1180	0.1548	0.1880	0.2674	0.1036	0.1459	0.14178	0.2069
ψ_{11}	0.0051	0.0072	0.0066	0.0093	0.0043	0.0068	0.00563	0.0072
ψ_{01}	-0.0165	-0.0252	-0.0090	-0.0236	-0.0132	-0.0232	-0.00132	-0.0108

*** $p < 0.0001$, ** $p < 0.001$, * $p < 0.01$, ~ $p < 0.05$

We begin by focusing upon differences in the estimates of fixed effects. The intercept term ($\hat{\beta}_{00}$) in each model is very close to 0, the expected mean scale score for third grade students in all eight scales. There are considerable differences in the average growth trajectory ($\hat{\beta}_{01}$) across models. In an absolute sense, growth appears much larger for scales in which calibration was based upon a combination of 3PLM/GPCM instead of 1PLM/PCM. This is consistent with the results we observed in Table 3—use of the 3PLM/GPCM will tend to stretch the range of the vertical score scale relative to the use of the 1PLM/PCM. Growth trajectories are always just slightly larger when scale scores are estimated using EAPs instead of MLEs. An important difference can be seen when comparing the scales created using the separate or hybrid approaches in conjunction with the 3PLM/GPCM. Whether scale scores are estimated using EAPs or MLEs, the growth trajectory under the separate approach is roughly .07 logits higher than under the hybrid approach.

Table 5. Decomposition of Variance and Correlations between Random Effects by Scale

	Separate Calibration				Hybrid Calibration			
	1PLM/PCM		3PLM/GPCM		1PLM/PCM		3PLM/GPCM	
	EAP	ML	EAP	ML	EAP	ML	EAP	ML
Total Variance	0.5895	0.9166	0.9367	1.5276	0.5158	0.8740	0.7295	1.2423
Variance Decomposition (Percentage of Total)								
Level 1 σ^2	12.0%	11.8%	13.2%	12.6%	11.9%	11.7%	12.6%	13.0%
Level 2 τ_{00}	65.6%	66.0%	65.0%	66.0%	65.7%	66.0%	65.1%	65.6%
Level 2 τ_{11}	1.5%	7.1%	1.6%	8.5%	1.3%	7.1%	2.6%	8.7%
Level 3 ψ_{00}	20.0%	16.9%	20.1%	17.5%	20.1%	16.7%	19.4%	16.7%
Level 3 ψ_{11}	0.9%	0.8%	0.7%	0.6%	0.8%	0.8%	0.8%	0.6%
Correlations (Within Levels)								
Student-level								
(τ00, τ11)	-0.84	-0.84	-0.21	-0.60	-0.76	-0.82	0.04	-0.46
School-level								
(ψ00, ψ11)	-0.67	-0.75	-0.26	-0.47	-0.63	-0.74	-0.05	-0.28

We can gain some insights into these differences by looking more closely at the estimates for the random effect variance components across models. A first point to note is that, not surprisingly, the models with the smallest average growth trajectories are the ones with the smallest amount of total variability, where total variability is the sum of level 1, 2 and 3 variance components. In Table 5 the variance component(s) at each level of the model is/are expressed as a percentage of the total. As a percentage of variance, there are really only two that show much movement across models: the student-level growth trajectory ($\hat{\tau}_{11}$), and the school-level intercept ($\hat{\psi}_{00}$). Interestingly, these terms move in opposite directions as a function of scale score estimation method, which has a substantial impact on the percentage of total variance due to differences in student growth. Another noticeable difference across models is the estimated correlation between student and school-level random effects. Under the 1PLM/PCM combination, this correlation is relatively strong and negative, regardless of whether EAPs or MLEs have been employed for the underlying scale. In contrast, under the 3PLM/GPCM combination, when EAPs are the basis for the underlying scale, the correlations drop dramatically.

A last comparison to consider is the empirical Bayes (EB) estimates of student and school random effects. These are often extracted from growth models as value-added estimates with a normative interpretation. To what extent to these estimates lead to similar conclusion about students and schools as a function of the underlying score scale? Table 6 provides the correlation matrix of student and school-level EB slope estimates.

Table 6. Correlations of Empirical Bayes Slope Estimates: Student and School-level

	sep1.eap	sep1.mle	sep3.eap	sep3.mle	hyb1.eap	hyb1.mle	hyb3.eap	hyb3.mle
sep1.eap		0.962	0.843	0.917	0.995	0.963	0.661	0.789
sep1.mle	0.959		0.732	0.901	0.936	0.999	0.525	0.756
sep3.eap	0.851	0.733		0.921	0.877	0.747	0.955	0.934
sep3.mle	0.923	0.904	0.922		0.917	0.912	0.806	0.953
hyb1.eap	0.995	0.932	0.884	0.921			0.715	0.807
hyb1.mle	0.961	0.999	0.748	0.914	0.936		0.548	0.775
hyb3.eap	0.685	0.538	0.960	0.811	0.736	0.560		0.903
hyb3.mle	0.809	0.767	0.939	0.957	0.824	0.785	0.905	

Note: Value above diagonal = EB slopes student-level; values below = EB slopes school-level

While all the random effect estimates are positively correlated across score scales, the strength of the correlation ranges considerably. The weakest correlations are found when scales calibrated under a hybrid approach using the 3PLM/GPCM combination with EAP estimation (“hyb3.eap”) are compared with scales calibrated under a separate approach using the 1PLM/PCM combination with MLE estimation (“sep1.mle”). All else held constant, the “main effect” choice of calibration approach, IRT model, and score estimation methodology seem to have little impact when the outcome of interest consists solely normative comparisons of student or school-level random effects. It is the interaction of these choices that have the greatest impact.

Discussion

Using longitudinal growth in student achievement as the basis for extracting information about 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 in growth and 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, in press; Raudenbush, 2004; Rubin, Stuart & Zanatto, 2004) One key assumption that has been often overlooked rests upon the way that student achievement is being measured and vertically scaled.

The state of Colorado currently places students scores on its CSAP tests onto a vertical scale. This scale is based upon a common-item nonequivalent groups design and makes use of a combination of 3PLM/GPCM IRT models, separate calibration and EAP estimation. The findings in this study suggest that had Colorado decided, for example, to instead use the entirely defensible combination of 1PLM/PCM models with hybrid calibration and ML estimation, the use of their vertical scale could lead to strikingly different educational accountability conclusions. In particular, the variability of scores along a vertical scale is very sensitive to the way the scale has been created. This can be problematic when change along the scale is given an absolute or criterion-based interpretation. Hence it would seem that states considering the application of growth to standard models should be especially cognizant of psychometric decisions being made in establishing their vertical scales. These seemingly esoteric decisions appear to have potentially substantial impacts on students and schools.

References

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

Carlson, D. (2006). *Focusing state educational accountability systems: 4 methods of judging quality and progress*. Retrieved November 21, 2006, from http://www.nciea.org/cgi-bin/pubspage.cgi.

CTB/McGraw-Hill. (2005). Colorado Student Assessment Program: 2005 technical report. Monterey, CA: CTB/ McGraw-Hill. Available online at http://www.cde.state.co.us/cdeassess/csap/as_technical.htm.

Goldschmidt, P., Roschewski, P., Choi, K. C., Auty, W., Hebbler, S., Blank, & Williams, A. (2005). *Policymakers' guide to growth models for school accountability: How do accountability models differ?* Washington, DC: CCSSO

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.

Hill, R., Gong, B., Marion, S., DePascale, C., Dunn, J., & Simpson, M. A. (2006). Using value tables to explicitly value growth. In R. Lissitz (Ed.), *Longitudinal and value added models of student performance* (pp. 255-290). Maple Grove, MN: JAM Press

Karkee, T., Lewis, D. , Hoskens, M.,Yao, L, and Haug, C. (2003) Separate versus concurrent calibration methods in vertical scaling. CTB Research Report. (Available online from www.eric.ed.gov)

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, 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. (2004) Linking assessments: concepts and history. *Applied Psychological Measurement*. 28(4), 219-226.

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

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

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

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*.

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

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

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, 159-176.

Pinheiro, J. C. & D. M Bates. 2000. Mixed-effects models in S and S-PLUS. New York: Springer.

Raudenbush, S. W. (2004a). *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

Raudenbush, S.W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edition. Sage Publications.

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.

Schmidt, W. (2004) The role of content in value-added. Presented at the conference *Value-Added Modeling: Issues with Theory and Application*, October 21, 2004, University of Maryland.

Singer, J. D. 1998. "Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models." Journal of Educational and Behavioral Statistics 24:323—355.

Spellings, M. (November 18, 2005). Secretary Spellings Announces Growth Model Pilot, Address Chief State School Officers' Annual Policy Forum in Richmond. *U.S. Department of Education Press Release*. Retrieved August 7, 2006 from http://www.ed.gov/news/pressreleases/2005/11/1182005.html.

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

Tong, Y, and Kolen, M. (2005) Comparison of methodologies and results in vertical scaling of educational achievement test. Paper presented at the annual meeting of the National Council for Measurement in Education, Montreal, CA.

U.S. Department of Education (2006, January 25). *Peer review guidance for the NCLB Growth model Pilot Applications.* Retrieved November 21, 2006 from http://www.ed.gov/admins/lead/account/growthmodel/index.html

Van der Linden, W. J., and Hambleton, R. K. (Eds.) (1997) *Handbook of modern item response theory*. New York: Springer-Verlag.

Venaples, W.N., and Smith, D.M. (2004) *An introduction to R*. Available online at http://cran.r-project.org/doc/manuals/R-intro.pdf

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

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