November 16, 2004

Dear CCSSO colleagues,

This is the paper containing the graphics I showed during the CCSSO Brain Trust on using student-level data for growth models in accountability.

This article has a technical methods section, but everything before and after the methods section (except two equations in the results section) is written so that it should be accessible to readers without any need to understand the mathematics in the methods section. The mathematics of the two equations in the results section are not necessary to understand, so long as the explanation in the text is understandable.

This article came from my dissertation, which goes beyond just value-added modeling, to show how the problems of changing content from one grade level of the test to another also cause serious problems for studies of the effectiveness of educational interventions, the effects of demographic variables, and with the estimation of growth trajectories in general. If anyone is interested in the larger document (which is much more technical), please e-mail me at martineauj@michigan.gov.

Thanks!

-- Joseph
Distorting Value Added: The Use of Longitudinal, Vertically Scaled Student Achievement Data for Growth-Based Value-Added Accountability

Joseph A. Martineau
Michigan Department of Education

Correspondence concerning this article should be addressed to

Joseph A. Martineau
Psychometrician
Office of Educational Assessment & Accountability
Michigan Department of Education
P.O. Box 30008
Lansing, MI 48909

(517) 241-4710 (Work)
(517) 410-5220 (Mobile)
(517) 335-1186 (Fax)

martineauj@michigan.gov

Acknowledgments

Professors Mark D. Reckase, Richard T. Houang, and Kenneth A. Frank of the Michigan State University Department of Counseling, Educational Psychology, and Special Education, and co-Director David N. Plank of the Michigan State University Educational Policy Center contributed significantly to the development of my dissertation, which provided the basis for this article. Three anonymous reviewers also provided valuable feedback, improving this article immensely.
ABSTRACT

Longitudinal, student performance-based value-added accountability models have become popular of late, and continue to enjoy increasing popularity. Such models require student data to be vertically scaled across wide grade and developmental ranges so that the value added to student growth/achievement by teachers, schools, and districts may be modeled in an accurate manner. Many assessment companies provide such vertical scales and claim that those scales are adequate for longitudinal value-added modeling. However, psychometricians tend to agree that scales spanning wide grade/developmental ranges also span wide content ranges, and that scores cannot be considered exchangeable along the various portions of the scale. This shift in the constructs being measured from grade to grade jeopardizes the validity of inferences made from longitudinal value-added models. This study demonstrates mathematically that the use of such "construct-shifting" vertical scales in longitudinal, value-added models introduces remarkable distortions in the value-added estimates of the majority of educators. These distortions include (1) identification of effective teachers/schools as ineffective (and vice versa) simply because their students' achievement is outside the developmental range measured well by "appropriate" grade-level tests, and (2) the attribution of prior teacher/school effects to later teachers/schools. Therefore, theories, models, policies, rewards, and sanctions based upon such value-added estimates are likely to be invalid because of distorted conclusions about educator effectiveness in eliciting student growth. This study identifies one valid use of student achievement data for growth-based value-added modeling: the assessment of an upper grade content (e.g. fourth grade) in both the grade below and the appropriate grade to obtain a measure of student gain on a grade-specific mix of constructs.
Distorting Value Added: The Use of Longitudinal, Vertically Scaled Student Achievement Data for Growth-Based Value-Added Accountability

LITERATURE REVIEW

Student-Performance-Based Accountability and Value-Added Assessment

The passage of the No Child Left Behind Act ("Nclb", 2002) legislates that individual states implement accountability systems based on student test scores to track Adequate Yearly Progress (AYP) and the closure of achievement gaps. For this and other reasons, the use of student test scores for accountability purposes is widespread (Goertz, Duffy, & Le Floch, 2001; Millman, 1997). A typical approach to achievement-based accountability is to track AYP and the closure of achievement gaps without tracking student cohorts (see Goertz et al., 2001).

This accountability use of cross-sectional data on successive cohorts is criticized as unfair to educators because it holds educators accountable for both student background, prior educational experience, and current educational effectiveness (Baker, Linn, Herman, & Koretz, 2002; Fuhrman & Elmore, 2004; Millman, 1997; Sanders & Horn, 1994; Thum, 2002).

Value-added assessment (VAA) is often suggested as a desirable alternative that holds educators accountable only for certain types of gains that students make during the time those educators’ taught their students (for examples of and discussions of this trend toward tracking student growth rates [or student gains toward a goal], see Baker et al., 2002; Fuhrman & Elmore, 2004; Linn, 2001; Millman, 1997; Sanders & Horn, 1994; Schacter, 2001; Thum, 2003; Webster, Mendro, & Almaguer, 1994; Westat & Policy Studies Associates, 2001). Finally, value-added (VA) approaches are increasingly entering the educational decision-making process (Baker et al., 2002; Herman, Brown, & Baker, 2000; Michigan State University Education Policy Center, 2002; Olson, 2002).

The Measurement Invariance Requirement of VAA Models

One critical requirement of VAA models is that longitudinal student achievement data used in the models must be based on vertically “equated” developmental scales which measure the same constructs across all grade levels of the assessment (Bryk, Thum, Easton, & Luppescu, 1998; Lewis, 2001; Linn, 2001; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Sanders & Horn, 1994; Thum, 2003).

However, in their jointly developed Standards for Educational and Psychological Testing (1985), the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) reserve the term equating for instruments of similar difficulty measuring the same underlying constructs, preferring to call the process of scaling instruments of differing content and difficulty scaling to achieve comparability (see Barnard, 1996 for a discussion of these terms).

Linn (1993) and Mislevy (1992) both explicitly identify vertical linking as calibration, or one step down the linking hierarchy from equating, where the scores can be considered imperfectly exchangeable. However, shortly after categorizing vertical linking as calibration, Linn (1993) implies that vertical scaling may fall another level down the hierarchy by remarking that "the calibration requirement that two tests measure the same thing is generally only crudely
approximated with tests designed to measure achievement at different developmental levels” (p. 91).

Mislevy (1992) also suggests that vertical scaling may be a weaker linkage than calibration by defining projection (“the weakest statistically based linkage procedure”) as concerning "assessments constructed around…the same [constructs] but with tasks that differ in format or content” (p. 62). As vertically scaled assessment batteries are purposely constructed around the same constructs, but differ in content from grade to grade, this description of projection matches that of vertical scaling. Mislevy (1992) states "Projection sounds rather precarious, and it is. The more assessments arouse different aspects of students' knowledge, skills, and attitudes, the wider the door opens for students to perform differently in different settings" (p. 63). Second, he states that in projection "We can neither equate [the two measures of different constructs] nor calibrate them to a common frame of reference, but we may be able to gather data about the joint distribution of scores among relevant groups of students." (p. 54).

Finally, Yen (1986) provides a compelling argument for not using vertically scaled data for modeling accountability, stating that the kinds of comparisons that studies of accountability are likely to include are not meaningful in this context:

It is also important to be more cautious in comparing results from tests that differ a great deal in difficulty or content. As a convenience, test publishers produce scales that go from Kindergarten to 12th grade. Although the legitimate purpose of such scales is to define correspondence between successive pairs of test levels, the existence of such a broad scale might lead test users to the illegitimate comparison of widely separated levels (p. 322).

The Need for This Study

The use of vertically-scaled student achievement data for growth-based accountability measures is increasing in use, and is suggested as fairer than non-cohort analysis. The APA, AERA, and NCME (1999) jointly state that when specific misinterpretations are likely, they should be explained to test users; that when non-parallel forms of an assessment are equated, the adequacy of the equating should be detailed; that when growth or gains are being measured, the validity of inferences based on those scores should be documented; and that when assessment data are used for high-stakes purposes, the validity of those uses should be subjected to heightened scrutiny. Because accountability use of growth-based VAA models is extraordinarily high-stakes, it is vital to understand the potential impact of violating the assumption of measurement invariance that arguably exists with vertical scales.

Research Question

This study provides a mathematical basis for determining the effects of violating the measurement invariance requirement on the interpretations derived from growth-based VAA models using vertically-scaled developmental scores. For ease in discussion, the term "construct shift" is coined to describe this violation. This study addresses the question: How do varying degrees and types of construct shift distort results of growth-based VAA models?
Desirable Interpretations of Value-Added Estimates

There are at least four reasonable types of interpretations that policy-makers may want to make using the VA estimates resulting from VAA models. It is against these four interpretations that distortions in results of growth-based VAA models are compared.

The first type of interpretation is the value units add to student gains on a single construct. This may be a desirable interpretation because it provides a “pure” measure of the effectiveness of a unit in teaching a single construct. However, this interpretation may be problematic because it assumes the there is a pure measure of each important construct being taught.

The second type of interpretation is the value units add to student gains on a static mix of constructs. This may be a desirable interpretation because it provides a combined measure of effectiveness in teaching multiple constructs (e.g. computation and problem solving in mathematics). However, this interpretation may be problematic because it assumes that emphases in the curriculum do not change over the grade levels included in the VA analysis.

The third type of interpretation is the value units add to student gains on a grade-specific mix of constructs where the mix is defined by the representation of the various constructs in grade-specific assessments. This may be a desirable interpretation because it provides a grade-specific combined measure of effectiveness in teaching multiple constructs, allowing for changes in construct emphases across grade levels. If the policy decision is to hold all units accountable for student growth on constructs defined by the curriculum and mirrored by the assessments, this is a reasonable interpretation. However, this may be problematic if the policy decision is to measure growth using the level of the test that best matches the mix of constructs where the student is primarily growing.

The fourth type of interpretation is the value units add to student gains on a student-tailored mix of constructs where the mix is defined by the best match of the test level to the developmental level of the individual student. This may be a desirable interpretation because it provides a combined measure of effectiveness in teaching mixes of multiple constructs that are tailored to the developmental level of each student. It provides for a measure of effectiveness in the constructs on which students are making their primary growth. This is a particularly attractive interpretation for units teaching student whose incoming developmental level is far above or below that specified in the grade-specific curriculum. However, this interpretation may be problematic if the policy decision is to hold units accountable for student gains on the construct mix present on the tests at the students’ grade levels.

METHODS

Simplifying Assumptions of This Study

In order to facilitate the mathematics of this article, several simplifying assumptions are made, which if relaxed would only increase the complexity of the effects of construct shift:

1. Only one subject is analyzed at a time.
2. Measurement occasion is cross-nested within student and within one type of organizational unit (e.g. classroom, grade in a school, or grade in a district).
3. No covariates are entered into the model, as in TVAAS (Sanders, Saxon, & Horn, 1997), nor are they needed.
4. Every student is tested in every grade and advances after the end of each grade.
5. The sample of units is stable across time, and students do not move in or out of any unit during the school year.

Defining Dimensionality

*Pure unidimensionality* is defined as measurement of the same, single construct at each grade level; *empirical multidimensionality*\(^{1}\) as measurement of the same set and mix of constructs at each grade level; and *empirical multidimensionality* as measurement of a changing set and/or mix of constructs at each grade level.

Defining Value Added

*No-Effects VA* assumes that each unit (e.g. teacher/classroom, grade within school, or grade within district) adds exactly the same value to every student’s gains. Therefore no unit effects on student gains appear in no-effects models. *Layered-Effects VA* assumes that each unit adds value to student gains only during the grade that the unit serves those students (Sanders et al., 1997).

Defining Purely Unidimensional True Scores

*Layered-Effects Purely Unidimensional True Score*. The layered-effects purely unidimensional definition of true score for a single student in a single construct is:

\[
l_{ij}^{lu} = t_j + \sum_{m=0}^{i} (g_{mj} + a_{k(m,j)}),
\]

where

- \(i, m\) indicate grade, with 0 and \(I\) being the lowest and highest grade, respectively;
- \(j\) indicates student;
- \(k\) indicates unit (e.g. teacher/classroom, grade in a school, grade in a district);
- \(t_{ij}^{lu}\) is student \(j\)'s layered-effects purely unidimensional (lu) true score at the end of grade \(i\);
- \(t_j\) is the true score of student \(j\) just before entering the lowest grade in the analysis;
- \(g_{mj}\) is the "natural gain" of student \(j\) during grade \(m\), or the mean gain student \(j\) would make in all grade-\(m\) units in the analysis; and
- \(a_{k(m,j)}\) is the "value added" to student \(j\)'s gain during grade \(m\) by student \(j\)'s grade-\(m\) unit \((k)\)

\((a_{k'(m,j)} \equiv 0\) for a grade-\(m\) unit \(k'\) of average effectiveness for student \(j)\).

\(^{1}\)This differs from the traditional definition of essential unidimensionality (see Stout, 2002) in two ways: (1) the mix of constructs is not defined as the dominant dimensions to be measured by the test, but as each dimension being important in its own right; and (2) the definition involves a longitudinal component.
No-Effects Purely Unidimensional True Score. Because in no-effects definitions, all units have the same effect on all students, the “value-added” expression $a_k(m,j)$ from the previous definition resolves to zero. Therefore, the No-Effects Purely Unidimensional True Score definition is:

$$t_{ij}^{nu} = t_j + \sum_{m=0}^{i} g_{mj},$$

where $t_{ij}^{nu}$ is student $j$'s no-effects purely unidimensional (nu) true score at the end of grade $i$.

These definitions of true score provide for two unique expectations of true score for a given student: one in which every unit is equally effective in eliciting growth for every student, and one in which the average effectiveness in each unit is unique, and the effectiveness of each unit is unique for each student. Thus, the same student may have two distinct expectations.

Defining Empirically Unidimensional True Scores

Specifying a method of mixing true scores on various pure constructs allows for a definition of empirically unidimensional true scores. Linear combinations of pure constructs are used here, but if non-linear combinations were used, the results would be more complex.

The linear combination method used here is proportions that sum to one, multiplied by true scores on the various pure constructs (for a discussion of why such an approach is needed, see Koretz, McCaffrey, & Hamilton, 2001). This definition allows for a convenient interpretation of the single true score: each pure construct accounts for the proportion of the single true score specified in the definition.

Layered-Effects Empirically Unidimensional True Score. The layered-effects empirically unidimensional definition of true score is:

$$t_{ij}^{le} = \sum_{c=1}^{C} p_c \left( t_{cj} + \sum_{m=0}^{i} \left( g_{cmj} + a_{ck(m,j)} \right) \right)$$

under the constraint $\sum_{c=1}^{C} p_c = 1$,

where

$t_{ij}^{le}$ is the layered-effects empirically unidimensional (le) combined true score of student $j$ at the end of grade $i$;

c indicates construct;

$C$ is the number of constructs that combine to make up the single true score;

$p_c$ is the proportion of the combined true score that is accounted for by construct $c$;

$t_{cij}$ is student $j$'s true score on construct $c$ at the end of grade $i$;

$t_{cj}$ is student $j$'s true score on construct $c$ just before entering the lowest grade in the analysis;

$g_{cmj}$ is student $j$’s "natural gain" on construct $c$ during grade $m$, or the gain student $j$ makes in construct $c$ during grade $m$ in a classroom of average effectiveness for student $j$;
\( a_{ck(m,j)} \) is the value-added to student \( j \)'s gains on construct \( c \) during grade \( m \) by the unit \( (k) \) student \( j \) attended during grade \( m \); and all other terms are as defined previously.

**No-Effects Empirically Unidimensional True Score.** Because the “value-added” terms resolve to zero for no-effects models, the no-effects, empirically unidimensional definition is:

\[
t^{ne}_{ij} = \sum_{c=1}^{C} p_c \left( t_{cj} + \sum_{m=0}^{i} g_{cmj} \right),
\]

where \( t^{ne}_{ij} \) is student \( j \)'s no-effects empirically unidimensional (\( ne \)), combined true score at the end of grade \( i \).

**Defining Empirically Multidimensional True Scores**

Allowing the proportions \( (p_c) \) to change across grade level (subscripting with an \( i \)) facilitates the definition of empirically multidimensional true scores. The proportions may change in any given that grade level proportions must sum to 1 (implying that changes in proportion from one grade to the next must sum to 0).

**Layered-Effects Empirically Multidimensional True Score.** The layered-effects, empirically multidimensional definition of true score is:

\[
t^{lm}_{ij} = \sum_{c=1}^{C} p_{ci} \left( t_{cj} + \sum_{m=0}^{i} g_{cmj} + \sum_{m=0}^{i} a_{ck(m,j)} \right) \text{ with }
\]

with the constraints

\[
\sum_{c=1}^{C} p_{ci} = 1, \sum_{c=1}^{C} (p_{ci} - p_{c(i-1)}) = \sum_{c=1}^{C} d_{ci} = 0, \text{ and } d_{c0} \equiv 0;
\]

where

\( t^{lm}_{ij} \) is the layered-effects empirically multidimensional (\( lm \)) combined true score of student \( j \) at the end of grade \( i \);

\( p_{ci} \) is the grade-level-\( i \) proportion of combined true scores accounted for by construct \( c \);

\( d_{ci} \) is the change in proportional representation of construct \( c \) in true scores from the end of grade \( i-1 \) to the end of grade \( i \) (with \( d_{c0} \equiv 0 \) because no data exists before grade 0).

**No-Effects, Empirically Multidimensional True Score.** Because the “value-added” terms resolve to zero for no-effects models, the no-effects, empirically multidimensional definition is

\[
t^{nm}_{ij} = \sum_{c=1}^{C} p_{ci} \left( t_{cj} + \sum_{m=0}^{i} g_{cmj} \right),
\]
where \( t_{ij}^{nm} \) is the no-effects (\( n \)), empirically multidimensional (\( m \)) combined true score of student \( j \) at the end of grade \( i \).

Statistical Accountability Models

Two types of accountability models are also derived, to mirror the no-effects and layered-effects value-added definitions (note that the statistical models are not applied to the corresponding true-score models). In these derivations, only population parameters enter the equations, so the equations derived here are the asymptotic results of VAA models rather than the results of any given application of a VAA model.

**Level-1 Model**

For both the no-effects and layered-effects accountability models, the level-1 model is

\[
y_{ij} = \sum_{m=0}^{i} \beta_{mj} + \beta_{ij} = \gamma_{ij} + \gamma_{(i-1)j} \quad \text{for} \quad i>0, \quad \text{and} \quad \beta_{0j} = \gamma_{0j} \quad \text{for} \quad i=0. \tag{8}
\]

This model is saturated (there are \( I+1 \) regression weights for \( I+1 \) grade-specific observations), resulting in predicted scores being equal to observed scores. \( \beta_{0j} \) is student \( j \)'s observed score at the end of grade 0, and \( \beta_{ij} \) (for \( i>0 \)) is student \( j \)'s observed gain from the end of grade \( (i-1) \) to the end of grade \( i \).

In these derivations, the \( y \)'s and \( \beta \)'s are general. To specify which true score model is being applied, they are superscripted with \( nu, lu, ne, le, nm, \) and \( lm \) as done above.

**Specification of the Level-2 Model**

For the no-effects and layered-effects models, respectively, the level-2 model is

\[
\beta_{nj} = \gamma_{i} + u_{ij} \quad \text{and} \quad \beta_{lj} = \gamma_{i} + u_{ij} + \tilde{a}_{ik} \tag{9}
\]

where for no-effects, the model reduces to an unconditional 2-level mixed model, and for layered-effects, the model reduces to an unconditional 2-level cross-nested model (see Raudenbush & Bryk, 2002). In these models

\( \beta_{nj} \) is student \( j \)'s no-effects observed gain in grade \( i \),

\( \beta_{lj} \) is student \( j \)'s layered-effects observed gain in grade \( i \),

\( \gamma_{i} \) is the mean "natural gain" in grade \( i \),

\( \tilde{a}_{ik} \) is unit \( k \)'s the mean effect on its students’ grade-\( i \) gains, and

\( u_{ij} \) is the deviation of student \( j \)'s observed grade-\( i \) gain in grade \( i \) from the mean "natural gain" in grade \( i \) plus the mean effect on student gains in grade \( i \) of the unit \( (k) \) attended by students in grade \( i \).
The meaning of $\tilde{a}_{ik}$ given here corresponds directly to the mean of $a_{k(i,j)}$ as described in the definitions of true scores. The important difference is that $\tilde{a}_{ik}$ is always estimated using scores on a single score scale.

Effects of Dimensionality on Estimates of Unit Effects

To obtain estimates of unit effects ($\tilde{a}_{ik}$), the no-effects expression $\beta_{ij}^{n*}$ is solved for $\gamma_i$:

$$\gamma_i = \beta_{ij}^{n*} - u_{ij}. \tag{10}$$

This result is then substituted into the layered-effects expression $\beta_{ij}^{l*}$, since $\gamma_i$ has the same meaning in both expressions. The resulting expression is then solved for $\tilde{a}_{ik}$. Therefore,

$$\beta_{ij}^{l*} = \beta_{ij}^{n*} - u_{ij} + u_{ij} + \tilde{a}_{ik} = \beta_{ij}^{l*} - \beta_{ij}^{n*}. \tag{11}$$

Because $\tilde{a}_{ik}$ varies at the unit level, the mean of both sides of the equation can be taken across students within unit, giving

$$\tilde{a}_{ik} = \mu_{\beta_{ik}^{l*}} - \mu_{\beta_{ij}^{n*}}, \tag{12}$$

where the $\mu$s are the unit-k means of the previously defined quantities.

Because from above, $\beta_{ij} = y_{ij} - y_{(i-1)j}$, and because the mean of measurement error within a given unit is assumed to be zero, the means of the $\beta$’s includes means of true scores rather than observed scores, or

$$\mu_{\beta_{ik}^{l*}} = \mu_{y_{ik}} - \mu_{y_{(i-1)k}} = \mu_{y_{ik}} - \mu_{t_{(i-1)k}} - \mu_{\epsilon_{(i-1)k}} = \mu_{\beta_{ik}^{l*}} - \mu_{t_{(i-1)k}} + \mu_{\epsilon_{(i-1)k}}, \tag{13}$$

where the $\mu$s are the unit-k means of the previously defined quantities.

The mean of measurement error within a classroom is assumed to be zero because (1) the definition of measurement error (from Classical Test Theory) includes the assumption that the mean of measurement error is zero by defining measurement error as the difference between an examinee’s observed score and his or her average of observed scores over an infinite number of repeated measures with the same assessment (see pages 109-110 of Crocker & Algina, 1986), and (2) this asymptotic definition of measurement error may be applied within units without changing the meaning.

The result in equation 13 is now substituted back into equation 12, giving

$$\tilde{a}_{ik} = \left( \mu_{f_{ik}^{*}} - \mu_{f_{(i-1)k}^{*}} \right) - \left( \mu_{n_{ik}^{*}} - \mu_{n_{(i-1)k}^{*}} \right) = \mu_{f_{ik}^{*}} - \mu_{f_{(i-1)k}^{*}} - \mu_{n_{ik}^{*}} + \mu_{n_{(i-1)k}^{*}}, \tag{14}$$

where the $\mu$s are the unit-k means of the previously defined quantities.
All that remains from this point is to calculate the mean of each expression of true score, and insert the appropriate mean true-score definitions into equation 14. The means, taken across persons within classroom are given in the Table 1 where the $\mu$ are the unit-$k$ grade-$i$ means of the previously defined quantities.

Table 1. Unit-mean true-score expressions.

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>True score model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No-effects</td>
</tr>
<tr>
<td>Purely unidimensional</td>
<td>$\mu_{nu} = \mu_k + \sum_{m=0}^{i} \mu_{gmk}$</td>
</tr>
<tr>
<td>Empirically unidimensional</td>
<td>$\mu_{ne} = \sum_{c=1}^{C} p_c (\mu_{tck} + \sum_{m=0}^{i} \mu_{gcmk})$</td>
</tr>
<tr>
<td>Empirically multidimensional</td>
<td>$\mu_{nm} = \sum_{c=1}^{C} p_{ci} (\mu_{tck} + \sum_{m=0}^{i} \mu_{gcmk})$</td>
</tr>
</tbody>
</table>

For clarity, the following symbols in Table 1 are defined explicitly:

- $\mu_{gmk}$ is the mean (across students in unit $k$ at grade $i$) of natural gains in grade-$m$,
- $\mu_{gcmk}$ is the mean (across students in unit $k$ at grade $i$) of construct-$c$ natural gains in grade-$m$,
- $\mu_{amk}$ is the mean (across students in unit $k$ at grade $i$) of value added to student gains by the units students attended in grade $m$, and
- $\mu_{acmk}$ is the mean (across students in unit $k$ at grade $i$) of value added to construct-$c$ student gains by the units students attended in grade $m$.

In the following derivations, these resulting means of true scores are inserted into equation 14. For brevity, intermediate steps of the algebraic manipulations are not shown.

### Purely Unidimensional Model

$$\bar{\alpha}_{ik} = \mu_{lu}^{i(i-1)k} - \mu_{lu}^{i(i-1)k} - \mu_{nu}^{i(i-1)k} + \mu_{nu}^{i(i-1)k} = \sum_{m=0}^{i} \mu_{amk} - \sum_{m=0}^{i-1} \mu_{amk} = \mu_{\bar{\alpha}_{ik}}$$  \hspace{1cm} (15)

### Empirically Unidimensional Model

$$\bar{\alpha}_{ik} = \mu_{le}^{i(i-1)k} - \mu_{le}^{i(i-1)k} - \mu_{ne}^{i(i-1)k} + \mu_{ne}^{i(i-1)k} = \sum_{c=1}^{C} p_c \left( \sum_{m=0}^{i} \mu_{acmk} - \sum_{m=0}^{i-1} \mu_{acmk} \right) = \sum_{c=1}^{C} p_c \mu_{\bar{\alpha}_{ik}} ,$$  \hspace{1cm} (16)

### Empirically Multidimensional Model

$$\bar{\alpha}_{ik} = \mu_{lm}^{i(i-1)k} - \mu_{lm}^{i(i-1)k} - \mu_{nm}^{i(i-1)k} + \mu_{nm}^{i(i-1)k} = \sum_{c=1}^{C} \left( p_{ci} \sum_{m=0}^{i} \mu_{acmk} - p_{c(i-1)} \sum_{m=0}^{i-1} \mu_{acmk} \right) ,$$  \hspace{1cm} (17)
\[ \tilde{a}_{ik} = \sum_{c=1}^{C} \left( p_{ci} \mu_{a_{cik}} + \left( p_{ci} - p_{c(i-1)} \right) \sum_{m=0}^{i-1} \mu_{a_{cik}} \right) = \sum_{c=1}^{C} p_{ci} \mu_{a_{cik}} + \sum_{c=1}^{C} \left( d_{ci} \sum_{m=0}^{i-1} \mu_{a_{cik}} \right), \] (18)

RESULTS

VAA Definitions for Scores of Varying Dimensionality

The final equations for the layered-effects VAA models are given in Table 2, where all symbols have been previously defined.

Table 2. VAA Definitions for Scores of Varying Dimensionality.

<table>
<thead>
<tr>
<th>Dimensionality of scores</th>
<th>Expression for value added by unit k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purely unidimensional</td>
<td>( \mu_{a_{ik}} )</td>
</tr>
<tr>
<td>Empirically unidimensional</td>
<td>( \sum_{c=1}^{C} p_{ci} \mu_{a_{cik}} )</td>
</tr>
<tr>
<td>Empirically multidimensional</td>
<td>( \sum_{c=1}^{C} p_{ci} \mu_{a_{cik}} + \sum_{c=1}^{C} \left( d_{ci} \sum_{m=0}^{i-1} \mu_{a_{cik}} \right) )</td>
</tr>
</tbody>
</table>

The Utility of VAA Results

The utility of the expressions in Table 2 are discussed in this section. Table 3 summarizes the support for the various desirable interpretations of VA provided by score scales of various dimensionality. The rationale for this summary is fully explained below the table.

Table 3. Support for value-added interpretations given by score scales of varying dimensionality.

<table>
<thead>
<tr>
<th>Desirable VA interpretation</th>
<th>Purely Unidimensional</th>
<th>Empirically Unidimensional</th>
<th>Empirically Multidimensional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single construct</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Static construct mix</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Grade-specific construct mix</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Student-tailored construct mix</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Purely Unidimensional Scores

In Table 2, the expression of value-added for a purely unidimensional score scale is exactly what one expects from a VAA model: the average effect of a unit on its students’ gains on a single construct. This expression supports the single-construct value-added interpretation, but no other. It is improbable that a purely unidimensional vertical score scale can be produced. Therefore, it is unlikely that VAA results based on a purely unidimensional score scale can be useful in practical settings.

Empirically Unidimensional Scores
In Table 2, the expression for an empirically unidimensional score scale also has an interpretation of interest: the weighted combination of a unit’s average effectiveness on the various constructs that combine to create the score scale, where weights reflect the constructs’ unchanging proportional representation in the single score scale. This expression supports the static construct mix value-added interpretation, but no other.

It is possible to construct a reasonably empirically unidimensional score scale by using carefully constructed and monitored parallel forms. The reasonableness and usefulness of this VAA estimate depends on the degree that the following assumptions hold:

1. the proportional construct representations on the score scale match the importance of the various constructs in the curriculum,
2. the importance of the various constructs in the curriculum does not change over the period of the VAA study (this implies that the grade span covered by the study is short enough that the importance does not change over time—say two testing cycles to cover one year),
3. the proportional construct representations on the score scale match the developmental level of the students taking the test, and
4. the score scale used is empirically unidimensional.

**Empirically Multidimensional Scores**

In Table 2, the expression for an empirically multidimensional score scale is has two terms. The first term of this expression, also has an interpretation of interest: the weighted combination of a unit’s average effectiveness on the various constructs that combine to create the score scale where weights reflect the constructs’ grade-specific representation in the single score scale. This particular term of the expression supports the grade-specific construct mix value-added interpretation, but no other.

If the first term were the only term in the expression, the reasonableness and usefulness of this VAA estimate would depend on the degree that the following assumptions hold:

1. the grade-specific proportional construct representations on the score scale match the grade-specific importance of the various constructs in the curriculum,
2. the grade-specific proportional construct representations on the score scale match the developmental level of the students taking each grade-specific form of the test.

However, the second term, or \[ \sum_{c=1}^{C} (d_{ci} \sum_{m=0}^{i-1} u_{c_{cmk}}) \], is a term that is never of interest in estimating a single unit’s value added: the weighted combination of all preceding units’ average effectiveness on the various constructs that combine to create the score scale where weights reflect the constructs’ grade-specific change in representation in the single score scale from the previous grade to the current grade. This term is never of interest because it distorts the value-added estimate of a single unit by contaminating the estimate with the effectiveness of other units.

---

2 The term *distortion* is used rather than *bias* because the expression is an unbiased estimate of a specific (less than useful) quantity, but gives a distorted expression for the value added by a single unit.
Parsing this expression is helpful in understanding the impact of construct shift on VAA measures. The weight \( (d_{ci}) \) is the change in proportional representation: where the proportional representation increases (or decreases), the sign of the weight is positive (or negative, respectively). This is multiplied by the mean value-added by all prior units, which is negative for low value added and positive for high value added. Therefore, units are benefited by being preceded by (1) units of high value added on constructs whose proportional representation increased from the previous grade level and (2) units of low value added on constructs whose proportional representation decreased from the previous grade level. Similarly, units are penalized for being preceded by (1) units of low value added on constructs whose proportional representation increased from the previous grade level and (2) units of high value added on constructs whose proportional representation decreased from the previous grade level.

It is useful to determine to which units the distortion (the second term) applies. The second term of the expression resolves to zero where (1) the changes in proportional representation and the mean prior-unit value-added on the various constructs combine fortuitously to cancel each other out, or (2) the mean prior-unit value-added is the same for all constructs. The first possibility is not discussed because it is unlikely and unobservable. The second possibility is always satisfied for the lowest grade in the analysis (because \( d_{c0} = 0 \)). For all other units, the distortions apply to some degree (the probability of having prior effectiveness exactly equal on all constructs is zero). Therefore, the only possible utility of the VAA estimates using empirically multidimensional scales is for units teaching the lowest grade in the analysis.

**Ameliorating Effects of Correlations among Constructs.** There is only one type of correlation among constructs that directly ameliorates distortions. Only where the correlations among units’ value added to student growth on the various constructs are very high (near unity) do concerns about distortions dissolve. All other types of correlations among constructs only indirectly ameliorate distortions by limiting the range of the correlations among units’ value added to growth on the various constructs to near unity. Therefore, the analysis of ameliorating effects of correlations among constructs is limited to correlations among units’ value added to student growth on the various constructs.

The practical effect of distortions and the ameliorating effects of inter-construct value-added correlations can be derived by comparing the magnitude of the distortion (the second term in the VAA expression for empirically multidimensional scale scores which never has an interesting interpretation) to the standard deviation of “the true combined value-added” (the first term in the VAA expression for empirically multidimensional scale scores, which does have an interpretation of interest), or

\[
\frac{\sum_{c=1}^{C} d_{ci} \sum_{m=0}^{i-1} \mu_{a_{cmk}}}{\sqrt{\text{var} \left( \sum_{c=1}^{C} P_{ci} \mu_{a_{cik}} \right)}} = \frac{\sum_{c=1}^{C} d_{ci} \sum_{m=0}^{i-1} \mu_{a_{cmk}}}{\sqrt{\sum_{c=1}^{C} \sum_{d=1}^{C} P_{ci} P_{di} \text{cov} \left( \mu_{ci}, \mu_{di} \right)}}
\]

(19)

Assuming that all pure-construct measures of value added are standardized, this simplifies to...
Because this expression contains the inter-construct value-added correlations in the denominator, the practical effects of distortions are indeed ameliorated to some degree by high inter-construct value-added correlations—the larger the correlations, the larger the denominator and the smaller the practical effect of the distortions. Even so, the concerns about distortions only dissolve to become negligible with inter-construct correlations very near unity because the correlations in this expression are under a radical.

An Hypothetical Application of VAA to an Empirically Multidimensional Score Scale

Effects of construct shift can be shown graphically using an hypothetical scenario in which an accountability model might be applied to an empirically multidimensional mathematics score scale. Four assumptions are made for ease of interpretation:

1. Overall math scores obtained from the grade level tests are comprised of only Basic Computation (BC) and Problem Solving (PS), plus measurement error.
2. Score scales of the two constructs (BC and PS) are equal-interval scales.
3. Construct shift occurs either linearly or non-linearly as in panels A or B of Figure 1.
4. The only observable score scale is the single score scale combining BC and PS.

In panel A of Figure 1, it is assumed that early grades' math scores are composed primarily of BC achievement, which drops off sharply in third grade in the case of non-linear construct shift, with PS assumed to increase sharply in proportional representation in fifth grade, and the construct mix becomes more settled toward later grades. In panel B, the shift in construct mix changes smoothly over time.

Figure 1. Non-linear and linear construct shift involving only two constructs.

Figures 2 and 3 show the results of a VAA study in which construct shift occurs as in panels A and B of Figure 1. In these figures, there are two sets of horizontal and vertical axes.
The first set of axes arranges the *panels* in the figures, representing the effectiveness of units in teaching BC and PS. The panels on the left and right represent units of low and high effectiveness in teaching BC, respectively. The panels on the bottom and top represent units of low and high effectiveness in teaching PS, respectively. Four of an infinite possible number of unit effectiveness profiles are represented in the figures (e.g. panels A represents units with a [high BC/low PS] effectiveness profile).

**Figure 2.** Example effects of linear construct shift on VAA estimates.

The second set of axes applies *inside* the panels. The horizontal axis of each panel is the grade level of the unit, and the vertical axis is the value of the combined effectiveness estimates of the units with the specified effectiveness profile for the panel. The scale on the vertical axis of each panel is standardized, where average effectiveness is defined as zero; and low and high effectiveness as –1 and 1, respectively, or one standard deviation above and below the mean.

In Figures 2 and 3, there are labeled gray lines in each panel. The gray lines represent the true BC, true PS, and true combined value-added of units in each panel. Figures 2 and 3 also have thin black lines marked with squares and diamonds. These represent units that were preceded by other units of varying effectiveness profiles. To make the presentation understandable, only a small selection of the possible prior effectiveness profiles is presented. The deviation of these lines from the line labeled “true combined VA” is the distortion of current units’ true VA attributable to the average VA by all units that previously taught the students in the current units (the second, distorted, term in the VAA expression for empirically multidimensional score scales). While it may seem intuitive that averaging VA across all prior units should make the prior VA profiles less variable, many students in smaller units come from a small number of previous units, and this assumption cannot be defended.
Figure 3. Example effects of non-linear construct shift on VAA estimates.

The impact of construct shift is obvious in these figures. The shape of the true combined VA curves is directly related to the shape of the construct-representation curves in Figure 1. Only where units are equally effective in eliciting growth in both constructs (the off diagonals of the figures) do the combined effectiveness estimates not follow the trajectories of construct representation shown in Figure 1. Where there are sharp changes in construct representation, there are sharp changes in combined effectiveness estimates even for units with exactly the same effectiveness profiles. Furthermore, in the main diagonals, the non-horizontal gray lines representing combined value added always lie between the undistorted value added for the individual constructs (BC and PS, represented by the horizontal gray lines). As seen in these figures, units that are effective in eliciting growth on the constructs heavily weighted by the appropriate grade level tests are identified as adding more value than other units.

This brings the discussion to the dark lines in Figures 2 and 3. Because the value-added estimates of units serving students in the lowest grade level included in the analysis are undistorted, units serving grades 3 in this scenario receive value-added estimates uncontaminated with prior units’ effectiveness. All other units represented by dark lines in Figures 2 and 3 are units whose value-added estimates are distorted by the inclusion of accumulated prior units’ value-added on the various constructs.

These figures show graphically that the distortion of prior units’ effectiveness is much stronger for units serving students at grade levels where the shift in construct mix is the largest. Particularly, in Figure 3 the units serving fifth grade students are likely to have larger distortions in their combined value-added estimates.

Figure 4 shows the practical effect of a distortion in VA that is attributable to a one-standard deviation difference between constructs in prior unit effectiveness (e.g. Ave BC/Low
PS or High PS/Ave BC), for both linear (panel A) and non-linear (panel B) construct shift, and for differing levels of correlation between constructs. This is done by applying equation 19 to this hypothetical scenario. Figure 4 does not include a line for a correlation of 1.0, since the concerns about construct shift disappear completely when effectiveness on the two constructs is perfectly correlated. Figure 4 also eliminates grade 3 because there is no effect of construct shift on the lowest grade in the analysis.

Figure 4. Practical effects of a one-standard deviation difference in effectiveness in BC/PS.

Figure 4 shows that the correlation of effectiveness in the various constructs has an impact on the practical effect of construct shift. With lower correlations, the practical effects are larger. For the linear construct shift detailed in Figure 1, the distortions in observed VA estimates range in size from 0.13 to 0.18 standard deviations of the true combined VA values. For the non-linear construct shift detailed in Figure 1, the distortions in observed VA estimates range in size from 0.02 to 0.49 standard deviations of the true combined VA values.

In essence, this shows that units’ value-added estimates may be consistently distorted by anywhere from 0 to nearly one half of a standard deviation, a particularly large level of uncertainty for use in high-stakes decisions.

DISCUSSION

Purely unidimensional score scales support a single-construct interpretation of value added to student gains. It is unlikely that one can create a vertical scale for a pure construct, and therefore VAA models are unlikely to be of practical utility for supporting an interpretation of value-added to a single construct.

Empirically multidimensional score scales support a static construct mix interpretation of value added to student gains. It is possible to create an empirically multidimensional score scale by carefully constructing parallel forms. VAA models built upon empirically multidimensional score scale may be of practical utility if the content of the assessments does not change with grade level or developmental level. This suggests a testing regime that measures fourth grade content in third grade and fourth grade, calculating value-added on that common scale; fifth grade content in fourth and fifth grade, calculating value added on that separate common scale, and so forth.

Empirically multidimensional score scales support a grade-specific construct mix interpretation of value-added to student gains, but only for units teaching the lowest grade in the
analysis. This is unlikely to be of practical utility as well. If multiple grades’ data is used in the analysis, it is only to obtain estimates of value added by units in the lowest grade. If only the lowest grade in the analysis is utilized, this leaves the door open to confounding of students’ incoming status with unit effects. Either approach is unsatisfactory.

The use of an empirically multidimensional score scale (e.g. a realistic use of vertical developmental scales) introduces distortions into the estimates of value-added by specific units by contaminating those estimates with the effectiveness of the units its students attended in prior years. This contamination is particularly strong where the change in content of the tests is greatest from one grade level to the next, to the point of identifying highly effective units as highly ineffective, or vice versa. Therefore, using empirically multidimensional score scales cannot be of practical use in measuring value-added by specific units.

This study has also shown that the intuitive hope that strong correlations among constructs will adequately alleviate the problems of construct shift is unfounded except in the most extraordinary circumstances: when units’ value-added to student growth in the various constructs are correlated nearly perfectly.

Therefore, with current technology, there are no vertical score scales that can be used to produce valid estimates of value added to student growth in either grade-specific or student-tailored construct mixes—the two most desirable interpretations of value added to student growth.

At this point, this leaves only one satisfactory approach to VAA using current technology: the measurement of a given grade-level’s content in both the grade below and the appropriate grade level to obtain an estimate of value-added to a static mix of constructs specific to each grade. This use of VAA is termed “paired-grade empirically unidimensional VAA.”

Implications, Limitations, and Future Research

It is established in this paper that (1) theoretical and applied psychometricians tend to agree that vertical scales shift in constructs being measured over grades, (2) value-added accountability models using construct shifted (empirically multidimensional) score scales are being put to extraordinarily high-stakes uses in a few circumstances with consideration to widen application, (3) the estimates resulting from such accountability systems are unmistakably distorted, and (4) a significant number of educators/schools are thus being unjustly identified as effective or ineffective. Thus, an admittedly serious, but reasonable implication of this study is to limit the high-stakes use of such value-added accountability systems to paired-grade empirically multidimensional applications, until VAA technology addresses the problems of construct shift identified in this study.

Although this study uses no data, the mathematical derivations are quite flexible, allowing for any reasonable scenario of student gains and teacher/school effectiveness. The extent to which single, combined scores are linear combinations of multiple construct scores is a concern, however. The single, combined scores may be non-linear combinations, and the results of this study would change to become more complex in describing the distortions in value-added estimates.

Further research needs to be performed to determine the degree to which construct shift actually occurs over the various grade levels of vertical scales. If value-added accountability
models are to be used on vertically scaled data, new methods of assuring that minimal construct shift occurs are imperative. These might include Multidimensional Item Response Theory methods of vertical scaling (Reckase, 1989, 1989, 1998; Reckase, Ackerman, & Carlson, 1988; Reckase & McKinley, 1991), or Multidimensional Computer Adaptive Testing (Luecht, 1996; Segall, 1996, 2000; van der Linden, 1999), keeping a close watch on when the various constructs enter into and leave the score scale; and reporting student scores on the various score scales within subject matter rather than reporting only a single combined "general math," "general reading," and/or "general science" scale.

If vertical scales can be developed that exhibit only minimal construct shift, then value-added accountability systems can be applied in additional ways that are valid and fair to educators (beyond paired-grade empirically unidimensional applications). If not, the distorted results of value-added accountability systems have the potential to cause great harm to the educational community, and little potential to work for the public good.

Finally, not considered in this paper is an additional requirement of VAA models that student scores exist on an equal-interval scale, and that whether this requirement is met is controversial (e.g., Angoff, 1971; Brennan, 1998; Camilli, 1999; Cliff, 1991; Schulz & Nicewander, 1997; Williams, Pommerich, & Thissen, 1998; Zwick, 1992). A future study should also investigate the effects of probable violations of this assumption of equal-interval scales on the effects of VAA models.

REFERENCES


