Journal of Educational and Behavioral Statistics
Spring 2006, Vol. 31, No. 1, pp. 35–62

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

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 (a) 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 (b) 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 highly restrictive scenarios in which current value-added models can be validly applied in high-stakes and low-stakes research uses. This article further identifies one use of student achievement data for growth-based, value-added modeling that is not plagued by the problems of construct shift: 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. Directions for future research on methods to alleviate the problems of construct shift are identified as well.

Keywords: *accountability, measurement of change, multidimensionality, school effects, teacher effects, validity, value-added models, vertical scaling*

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.

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 require-

ment 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). Because 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 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).

Yen (1986) also 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)

These four sources provide ample reason to question the utility of vertically scaled student achievement data for use in VAA.

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 noncohort analysis. The APA, AERA, and NCME (1999) jointly state that when specific misinterpretations are likely, they should be explained to test users; that when nonparallel 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 article 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 two questions: (a) *How do varying degrees and types of construct shift distort results of growth-based VAA models?* (b) *To what degree are the distortions associated with construct shift ameliorated when the different constructs are correlated?*

Desirable Interpretations of Value-Added Estimates

There are at least four reasonable types of interpretations about educational units (e.g., classrooms, grades within schools, or grades within districts) 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 students whose average 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

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-classified 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 (see Ballou, Sanders, & Wright, 2004 for an assertion that this is the case).
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 unidimensionality*[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:

$$t_{ij}^{lu} = t_j + \sum_{m=0}^{i} \left(g_{mj} + a_{k_{mj}} \right), \tag{1}$$

where

i, m = grade, with 0 and I being the lowest and highest grade, respectively;
j = student;

k_{mj} = the unit (e.g., teacher/classroom, grade in a school, grade in a district) that student j attended in grade m;

t_{ij}^{lu} = student j's layered-effects purely unidimensional (lu) true score at the end of grade i;

t_j = the true score of student j just before entering the lowest grade in the analysis;

g_{mj} = the "natural gain" of student j during grade m, or the mean gain student j would make across all grade-m units in the analysis; and

$a_{k_{mj}}$ = the "value added" to student gains by unit k_{mj} (with $a_{k'_{mj}} \equiv 0$ for a unit k'_{mj} of average effectiveness). Note that this value varies across units, not across students.

No-Effects Purely Unidimensional True Score

Because in no-effects definitions, all units have the same effect on student gains, the "value-added" expression $a_{k_{mj}}$ from the previous definition resolves to zero. Therefore, the No-Effects Purely Unidimensional True Score definition is:

$$t_{ij}^{nu} = t_j + \sum_{m=0}^{i} g_{mj}, \tag{2}$$

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 nonlinear 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:

$$t_{ij}^{le} = \sum_{c=1}^{C} p_c \left[t_{cj} + \sum_{m=0}^{i} \left(g_{cmj} + a_{ck_{mj}} \right) \right] \text{ under the constraint } \sum_{c=1}^{C} p_c = 1, \tag{3}$$

where

t_{ij}^{le} = the layered-effects empirically unidimensional (*le*) combined true score of student *j* at the end of grade *i;*
c = construct;
C = the number of constructs that combine to make up the single true score;
p_c = the proportion of the combined true score that is accounted for by construct *c;*
t_{cj} = student *j*'s true score on construct *c* just before entering the lowest grade in the analysis;
g_{cmj} = 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;*
$a_{ck_{mj}}$ = the value-added to student gains on construct *c* by unit k_{mj};

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:

$$t_{ij}^{ne} = \sum_{c=1}^{C} p_c \left(t_{cj} + \sum_{m=0}^{i} g_{cmj} \right), \tag{4}$$

where t_{ij}^{ne} 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 for student *j* in grade *i* is:

$$t_{ij}^{lm} = \sum_{c=1}^{C} p_{ci} \left(t_{cj} + \sum_{m=0}^{i} g_{cmj} + \sum_{m=0}^{i} a_{ck_{mj}} \right) \tag{5}$$

with the constraints

$$\sum_{c=1}^{C} p_{ci} = 1, \sum_{c=1}^{C} \left(p_{ci} - p_{c(i-1)} \right) = \sum_{c=1}^{C} d_{ci} = 0, \text{ and } d_{c0} \equiv 0; \tag{6}$$

where

t_{ij}^{lm} = the layered-effects empirically multidimensional (*lm*) combined true score of student *j* at the end of grade *i;*

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

d_{ci} = 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 exist before grade 0); and all other terms are as defined previously.

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

$$t_{ij}^{nm} = \sum_{c=1}^{C} p_{ci} \left(t_{cj} + \sum_{m=0}^{i} g_{cmj} \right), \tag{7}$$

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 derived independent of the corresponding true score models, meaning that they are not based on those 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

$$y_{ij} = \sum_{m=0}^{i} \beta_{mj}, \beta_{ij} = y_{ij} - y_{(i-1)j} \text{ for } i > 0, \text{ and } \beta_{0j} = y_{oj} \text{ for } i = 0 \tag{8}$$

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. β_{0j} is student j's observed score at the end of grade 0, and β_{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 ys and βs 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

$$\beta_{ij}^{n\bullet} = \gamma_i + u_{ij} \tag{9}$$

and

$$\beta_{ij}^{l\bullet} = \gamma_i + u_{ij} + a_{k_{\mathbf{ij}}}, \tag{10}$$

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-classified model (see Raudenbush & Bryk, 2002). In these models

$\beta_{ij}^{n\bullet}$ = student j's no-effects observed gain in grade i,
$\beta_{ij}^{l\bullet}$ = student j's layered-effects observed gain in grade i,
γ_i = the mean "natural gain" in grade i,
$k_{i\mathbf{j}}$ = the unit attended in grade i by the set of students ($\mathbf{j}$) that attended the same unit as student j in grade i,
$\hat{a}_{k_{i\mathbf{j}}}$ = unit $k_{i\mathbf{j}}$'s value added to student gains, and
u_{ij} = the deviation of student j's observed grade-i gain in grade i from the mean "natural gain" in grade i plus the mean value-added to student gains by unit $k_{i\mathbf{j}}$.

Note that the meaning of $\hat{a}_{k_{i\mathbf{j}}}$ given here corresponds directly to the meaning of $a_{k_{ij}}$ as described in the definitions of true scores.

Effects of Dimensionality on Estimates of Unit Effects

Subtracting Equation 9 from Equation 10,

$$\hat{a}_{k_{i\mathbf{j}}} = \beta_{ij}^{l\bullet} - \beta_{ij}^{n\bullet}. \tag{11}$$

Because $\hat{a}_{k_{i\mathbf{j}}}$ varies at the unit level (it is assumed to be the same for all students within a classroom), the mean of both sides of the equation can be taken across students within unit, giving

$$\hat{a}_{k_{i\mathbf{j}}} = m_{\mathbf{j}}\left(\beta_{ij}^{l\bullet}\right) - m_{\mathbf{j}}\left(\beta_{ij}^{n\bullet}\right) = m_{\mathbf{j}}\left(\beta_{ij}^{l\bullet} - \beta_{ij}^{n\bullet}\right), \tag{12}$$

where $m_{\mathbf{j}}$ indicates that the mean is taken across the set $\mathbf{j}$ of students.

It is assumed that observed scores are composed of true scores plus measurement errors, or,

$$y_{ij}^{\cdot\cdot\cdot} = t_{ij}^{\cdot\cdot\cdot} + \epsilon_{ij}^{\cdot\cdot\cdot}, \tag{13}$$

and that the mean of measurement error within a given unit and measurement occasion is zero (from Classical Test Theory: see pages 109–110 of Crocker & Algina, 1986). Because of this, and because $\beta_{ij}^{\cdot\cdot\cdot} = y_{ij}^{\cdot\cdot\cdot} - y_{(i-1)j}^{\cdot\cdot\cdot}$ from above, the means of the β's include only means of true scores rather than observed scores, or

$$\hat{a}_{k_{\mathbf{ij}}} = m_{\mathbf{j}}\left[\left(y_{ij}^{l\bullet} - y_{(i-1)j}^{l\bullet}\right) - \left(y_{ij}^{n\bullet} - y_{(i-1)j}^{n\bullet}\right)\right] = m_{\mathbf{j}}\left[t_{ij}^{l\bullet} - t_{ij}^{n\bullet} - \left(t_{(i-1)j}^{l\bullet} - t_{(i-1)j}^{n\bullet}\right)\right]. \tag{14}$$

The next step is to derive the expression for differences in true scores, and insert the appropriate mean true score definitions into Equation 14. For brevity, only the last steps in the derivations are provided here. For purely unidimensional data,

$$t_{ij}^{lu} - t_{ij}^{nu} - \left[t_{(i-1)j}^{lu} - t_{(i-1)j}^{nu}\right] = \sum_{m=0}^{i} a_{k_{mj}} - \sum_{m=0}^{i-1} a_{k_{mj}} = a_{k_{ij}}. \tag{15}$$

For empirically unidimensional data,

$$t_{ij}^{le} - t_{ij}^{ne} - \left[t_{(i-1)j}^{le} - t_{(i-1)j}^{ne}\right] = \sum_{c=1}^{C} p_c \left(\sum_{m=0}^{i} a_{ck_{mj}} - \sum_{m=0}^{i-1} a_{ck_{mj}}\right) = \sum_{c=1}^{C} p_c a_{ck_{ij}}. \tag{16}$$

Finally, for empirically multidimensional data,

$$t_{ij}^{lm} - t_{ij}^{nm} - \left[t_{(i-1)j}^{lm} - t_{(i-1)j}^{nm}\right] = \sum_{c=1}^{C}\left(p_{ci}\sum_{m=0}^{i} a_{ck_{mj}}\right) - \sum_{c=1}^{C}\left[p_{c(i-1)}\sum_{m=0}^{i-1} a_{ck_{mj}}\right] \tag{17}$$

$$t_{ij}^{lm} - t_{ij}^{nm} - \left[t_{(i-1)j}^{lm} - t_{(i-1)j}^{nm}\right] = \sum_{c=1}^{C}\left[p_{ci}a_{ck_{ij}} + \sum_{m=0}^{i-1} a_{ck_{mj}}\left(p_{ci} - p_{c(i-1)}\right)\right] \tag{18}$$

$$t_{ij}^{lm} - t_{ij}^{nm} - \left[t_{(i-1)j}^{lm} - t_{(i-1)j}^{nm}\right] = \sum_{c=1}^{C}\left(p_{ci}a_{ck_{ij}} + \sum_{m=0}^{i-1} d_{ci}a_{ck_{mj}}\right). \tag{19}$$

The final step is to take the means of these difference score expressions across the set **j** of students. For purely unidimensional data, the result is

$$\widehat{a}_{k_{i\mathbf{j}}} = m_{\mathbf{j}}\left(a_{k_{ij}}\right) = a_{k_{i\mathbf{j}}}, \tag{20}$$

because $a_{k_{ij}}$ does not vary over students.

For empirically unidimensional data, the result is

$$\widehat{a}_{k_{i\mathbf{j}}} = m_{\mathbf{j}}\left(\sum_{c=1}^{C} p_c a_{ck_{ij}}\right) = \sum_{c=1}^{C} p_c a_{ck_{i\mathbf{j}}}. \tag{21}$$

because $a_{ck_{ij}}$ does not vary over students.

Finally, for empirically multidimensional data, the result is

$$\widehat{a}_{k_{i\mathbf{j}}} = m_{\mathbf{j}}\left[\sum_{c=1}^{C}\left(p_{ci}a_{ck_{ij}} + \sum_{m=0}^{i-1} d_{ci}a_{ck_{mj}}\right)\right] = \sum_{c=1}^{C}\left(p_{ci}a_{ck_{i\mathbf{j}}} + \frac{1}{n_{\mathbf{j}}}\sum_{j=1}^{n_{\mathbf{j}}}\sum_{m=0}^{i-1} d_{ci}a_{ck_{mj}}\right), \tag{22}$$

or

$$\widehat{a}_{k_{i\mathbf{j}}} = \sum_{c=1}^{C} p_{ci} a_{ck_{i\mathbf{j}}} + \frac{1}{n_{\mathbf{j}}} \sum_{c=1}^{C} \sum_{j=1}^{n_{\mathbf{j}}} \sum_{m=0}^{i-1} d_{ci} a_{ck_{mj}}. \tag{23}$$

However, some prior units are likely to have been attended by more than one student from set **j**. Therefore, for empirically multidimensional data,

$$\widehat{a}_{k_{i\mathbf{j}}} = \sum_{c=1}^{C} p_{ci} a_{ck_{i\mathbf{j}}} + \frac{1}{n_{\mathbf{j}}} \sum_{c=1}^{C} \left(d_{ci} \sum_{k'=1}^{n'_{\mathbf{j}}} n_{\mathbf{j}k'} a_{ck'} \right). \tag{24}$$

where k' indexes a prior unit attended by at least one student in the set **j,** $n'_{\mathbf{j}}$ is the number of unique prior units attended by students in the set **j,** $n_{\mathbf{j}k'}$ is the number of students from the set **j** that attended prior unit k', and $a_{ck'}$ is the value added to student gains in construct c by prior unit k'.

Results

VAA Definitions for Scores of Varying Dimensionality

The final equations for the layered-effects VAA models are given in Table 1, where all symbols have been previously defined.

The Utility of VAA Results

Purely Unidimensional Scores

In Table 1, the expression of value-added for a purely unidimensional score scale is exactly what one expects from a VAA model: *the 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.

TABLE 1
VAA Definitions for Score Scales of Varying Dimensionality

Dimensionality of Scores	Expression for Value Added by Unit $k_{i\mathbf{j}}$
Purely unidimensional	$a_{k_{i\mathbf{j}}}$
Empirically unidimensional	$\sum_{c=1}^{C} p_c a_{ck_{i\mathbf{j}}}$
Empirically multidimensional	$\sum_{c=1}^{C} p_{ci} a_{ck_{i\mathbf{j}}} + \frac{1}{n_{\mathbf{j}}} \sum_{c=1}^{C} \left(d_{ci} \sum_{k'=1}^{n'_{\mathbf{j}}} n_{\mathbf{j}k'} a_{ck'} \right),\ d_{c0} \equiv 0$

Empirically Unidimensional Scores

In Table 1, the expression for an empirically unidimensional score scale also has an interpretation of interest: *the weighted combination of a unit's 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 1 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 1, the expression for an empirically multidimensional score scale is has two terms. The first term of this expression, or

$$\sum_{c=1}^{C} p_{ci} a_{ck_{ij}},$$

also has an interpretation of interest: *the weighted combination of a unit's value added 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

$$\frac{1}{n_{\mathbf{j}}}\sum_{c=1}^{C}\left(d_{ci}\sum_{k'=1}^{n'_{\mathbf{j}}}n_{\mathbf{j}k'}a_{ck'}\right),$$

is a term that is *never* of interest in estimating a single unit's value added: *the weighted combination of the accumulation of all preceding units' value-added 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, averaged across students in the unit.* This term is *never* of interest because it *distorts*[2] the value-added estimate of a single unit by contaminating the estimate with the effectiveness of other units.

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 accumulation of value-added by all prior units, averaged across students. This quantity is negative for low accumulations of value added and positive for high accumulations. Therefore, units are benefited by being preceded by (a) units of high value added on constructs whose proportional representation increased from the previous grade level and (b) units of low value added on constructs whose proportional representation decreased from the previous grade level. Similarly, units are penalized for being preceded by (a) units of low value added on constructs whose proportional representation increased from the previous grade level and (b) 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 (a) the changes in proportional representation and the accumulated prior unit value-added on the various constructs combine fortuitously to cancel each other out, or (b) 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 $d_{c0} \equiv 0$). 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 incontestable utility of the VAA estimates using empirically multi-dimensional 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 interconstruct correlations in value added to student growth are very high (near unity) do concerns about distortions disappear. All other types of interconstruct correlations only indirectly ameliorate distortions by limiting the range of interconstruct correlations in value added to near unity. Therefore, the analysis of ameliorating effects of correlations among constructs is limited to interconstruct correlations in value added to student growth.

Following the classical test theory definition of reliability as the ratio of the variance of construct-related differences to the variance of the sum of construct-related differences and irrelevant differences (see Crocker & Algina, 1986), the reliability of asymptotic estimates of empirically multidimensional value-added[3] can be calculated by

$$\frac{\text{var}\,(\textit{true combined VA})}{\text{var}\,(\textit{true combined VA} + \textit{distortion in VA})}, \tag{25}$$

or the ratio of the variance of the true combined value added to the variance of the asymptotic estimates of empirically multidimensional value added.

For simplicity of the derivations, it is assumed that (a) value added on all constructs at all grade levels is distributed standard normally, (b) the value added by one unit is independent of the value added by all other units,[4] and (c) interconstruct correlation of value added is the same across all grade levels.

Because of assumption (a), the true combined value added is distributed

$$\sim N\left[0,\ \text{var}\left(\sum_{c=1}^{C} p_{ci} a_{ck_{ij}}\right)\right], \tag{26}$$

where the variance of the true combined value added is

$$\sum_{c=1}^{C} p_{ci}^2 \sigma_{a_{ci}}^2 + 2\sum_{c=2}^{C}\sum_{c'=1}^{c-1} p_{ci} p_{c'i} \sigma_{a_{ci}, a_{c'i}}. \tag{27}$$

Because of assumption (c), the subscript i on the value added can be ignored. Because of assumption (a), the covariances are replaced with correlations, and the variances are replaced with ones. Therefore, the variance of true combined value added simplifies to

$$\sum_{c=1}^{C} p_{ci}^2 + 2\sum_{c=2}^{C}\sum_{c'=1}^{c-1} p_{ci} p_{c'i} \rho_{a_c,\, a_{c'}}. \tag{28}$$

If it is assumed that there are only two constructs that combine to create an overall score (and therefore $p_{2i} = 1 - p_{1i}$), then the variance further simplifies to

$$1 + 2p_{1i}^2 - 2p_{1i} + 2p_{1i}(1 - p_{1i})\rho_{a_1,a_2} = 1 - 2p_{1i}(1 - p_{1i})(1 - \rho_{a_1,a_2}). \tag{29}$$

Because of assumption (b) the variance of empirically multidimensional value added is simply the variance of true combined value added plus the variance of the distortion in empirically multidimensional value added, and the formula for the reliability becomes

$$\frac{\text{var}\,(\textit{true combined VA})}{\text{var}\,(\textit{true combined VA}) + \text{var}\,(\textit{distortion in VA})}. \tag{30}$$

Because of assumption (a), the distortion is distributed

$$\sim N\left\{0,\ \text{var}\left[\frac{1}{n_{\mathbf{j}}}\sum_{c=1}^{C}\left(d_{ci}\sum_{k'=1}^{n'_{\mathbf{j}}} n_{\mathbf{j}k'}a_{ck'}\right)\right]\right\}. \tag{31}$$

Because of assumptions (b) and (c) the variance can be calculated by summing the variances for each prior unit. Therefore, the variance becomes

$$n_{\mathbf{j}}^{-2}\sum_{k'=1}^{n'_{\mathbf{j}}}\text{var}\left[\sum_{c=1}^{C}\left(d_{ci}n_{\mathbf{j}k'}a_{ck'}\right)\right] = n_{\mathbf{j}}^{-2}\sum_{k'=1}^{n'_{\mathbf{j}}}\left(\sum_{c=1}^{C}d_{ci}^2 n_{\mathbf{j}k'}^2\sigma_{a_c}^2 + 2\sum_{c=1}^{C}\sum_{c'=1}^{C}d_{ci}d_{c'i}n_{\mathbf{j}k'}^2\sigma_{a_c,a_{c'}}^2\right), \tag{32}$$

or

$$n_{\mathbf{j}}^{-2}n'_{\mathbf{j}}\left(\sum_{c=1}^{C}d_{ci}^2\sigma_{a_c}^2 + 2\sum_{c=1}^{C}\sum_{c'=1}^{C}d_{ci}d_{c'i}\sigma_{a_c,a_{c'}}\right)\sum_{k'=1}^{n'_{\mathbf{j}}}n_{\mathbf{j}k'}^2. \tag{33}$$

Because of assumption (a), the variance further simplifies to

$$n_{\mathbf{j}}^{-2}n'_{\mathbf{j}}\left(\sum_{c=1}^{C}d_{ci}^2 + 2\sum_{c=1}^{C}\sum_{c'=1}^{C}d_{ci}d_{c'i}\rho_{a_c,a_{c'}}\right)\sum_{k'=1}^{n'_{\mathbf{j}}}n_{\mathbf{j}k'}^2. \tag{34}$$

If it is assumed that there are only two constructs that combine to create an overall score (and therefore $d_{2i} = -d_{1i}$), the variance becomes

$$n_{\mathbf{j}}^{-2}n'_{\mathbf{j}}\left(d_{1i}^2 + d_{2i}^2 + 2d_{1i}d_{2i}\rho_{a_1,a_2}\right)\sum_{k'=1}^{n'_{\mathbf{j}}}n_{\mathbf{j}k'}^2 = 2d_{1i}^2\left(1-\rho_{a_1,a_2}\right)n_{\mathbf{j}}^{-2}n'_{\mathbf{j}}\left(\sum_{k'=1}^{n'_{\mathbf{j}}}n_{\mathbf{j}k'}^2\right). \tag{35}$$

But, the minimum value of the expression

$$\sum_{k'=1}^{n'_{\mathbf{j}}}n_{\mathbf{j}k'}^2 \tag{36}$$

is the simultaneous solution of the $n'_{\mathbf{j}}$ first derivatives of the expression with respect to each $n_{\mathbf{j}k'}$ set equal to zero, or

$$\frac{d\left(\sum_{k'=1}^{n'_{\mathbf{j}}}n_{\mathbf{j}k'}^2\right)}{d\left(n_{\mathbf{j}k^*}\right)} = \left(\sum_{k'=1}^{n'_{\mathbf{j}}}n_{\mathbf{j}k'}^2\right) - n_{\mathbf{j}k^*}^2 + 2n_{\mathbf{j}k^*} = \left(\sum_{k'=1}^{n'_{\mathbf{j}}}n_{\mathbf{j}k'}^2\right) + n_{\mathbf{j}k^*}\left(2 - n_{\mathbf{j}k^*}\right) = 0. \tag{37}$$

where n_{jk^*} represents any $n_{jk'}$. Subtracting any of these expressions from any other,

$$n_{jk^*}\left(2 - n_{jk^*}\right) - n_{jk^{**}}\left(2 - n_{jk^{**}}\right) = 0 \tag{38}$$

for all $k^* \neq k^{**}$. This expression is true when $n_{jk^*} = n_{jk^{**}}$, or when all prior units were attended by the same number of students from the set **j.** Mathematically, this is expressed as

$$n_{jk'} = i n_{\mathbf{j}} n_{\mathbf{j}}'^{-1}. \tag{39}$$

for all k'. Therefore

$$\sum_{k'=1}^{n'_{\mathbf{j}}} n_{jk'}^2 \geq \sum_{k'=1}^{n'_{\mathbf{j}}} \left(i n_{\mathbf{j}} n_{\mathbf{j}}'^{-1}\right)^2 = \sum_{k'=1}^{n'_{\mathbf{j}}} i^2 n_{\mathbf{j}}^2 n_{\mathbf{j}}'^{-2} = n'_{\mathbf{j}} i^2 n_{\mathbf{j}}^2 n_{\mathbf{j}}'^{-2} = i^2 n_{\mathbf{j}}^2 n_{\mathbf{j}}'^{-1}. \tag{40}$$

Inserting this result into Equation 35, the minimum value of the variance of the distortion in VA can be expressed as

$$\text{var}\,(\textit{distortion}) \geq 2d_{1i}^2\left(1 - \rho_{a_1,a_2}\right) n_{\mathbf{j}}^{-2} n'_{\mathbf{j}} i^2 n_{\mathbf{j}}^2 n_{\mathbf{j}}'^{-1} = 2i^2 d_{1i}^2\left(1 - \rho_{a_1,a_2}\right). \tag{41}$$

Inserting Equations 29 and 41 into Equation 30, the upper bound on reliability of value-added estimates is

$$\frac{1 - 2p_{1i}\left(1 - p_{1i}\right)\left(1 - \rho_{a_1,a_2}\right)}{1 - 2p_{1i}\left(1 - p_{1i}\right)\left(1 - \rho_{a_1,a_2}\right) + 2i^2 d_{1i}^2\left(1 - \rho_{a_1,a_2}\right)}. \tag{42}$$

This equation makes it clear that when the interconstruct correlation of value added is perfect, the impact of construct shift disappears (the reliability of empirically multidimensional value added becomes one). It also makes it clear that larger changes in proportional representation result in lower reliability, and that construct shift results in the maximum value of the reliability being lower in later grades.

A 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 nonlinearly 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 Figure 1, it is assumed that early grades' math scores are composed primarily of BC achievement, with later grades' math scores being composed primarily of PS achievement. In Panel A, the shift in proportional representation occurs linearly over time. In Panel B, the proportional representations change sharply between fourth and fifth grade.

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, represent-

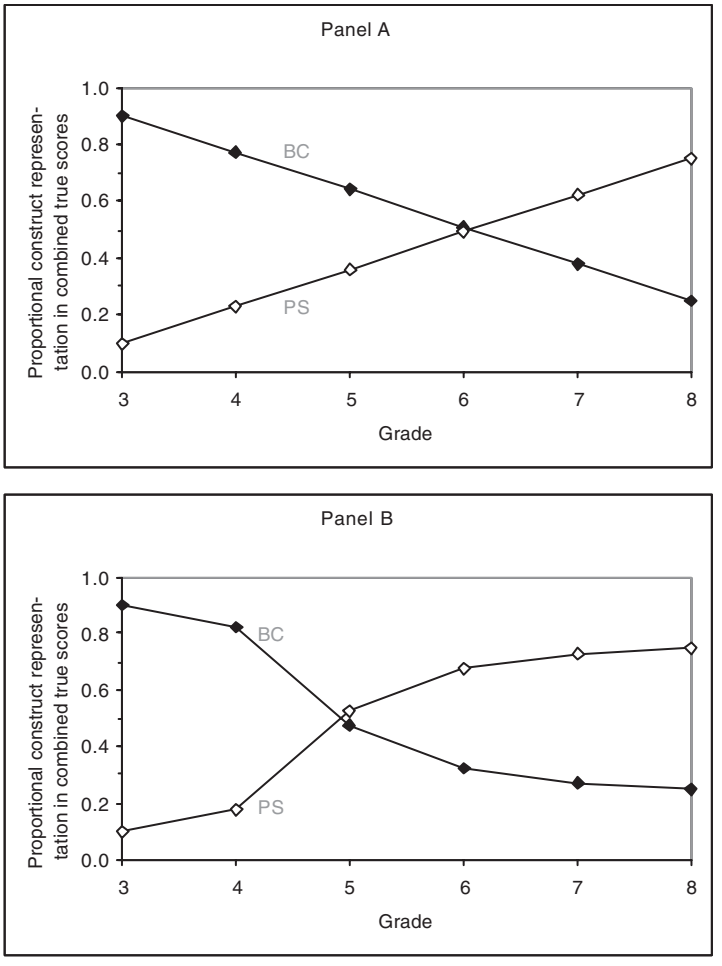

FIGURE 1. *Linear (**A**) and nonlinear (**B**) construct shift involving only two constructs.*

ing 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., panel A represents units with a [high BC/low PS] effectiveness profile).

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.

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 (panels B and C) 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 panels A and D, the nonhorizontal 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.

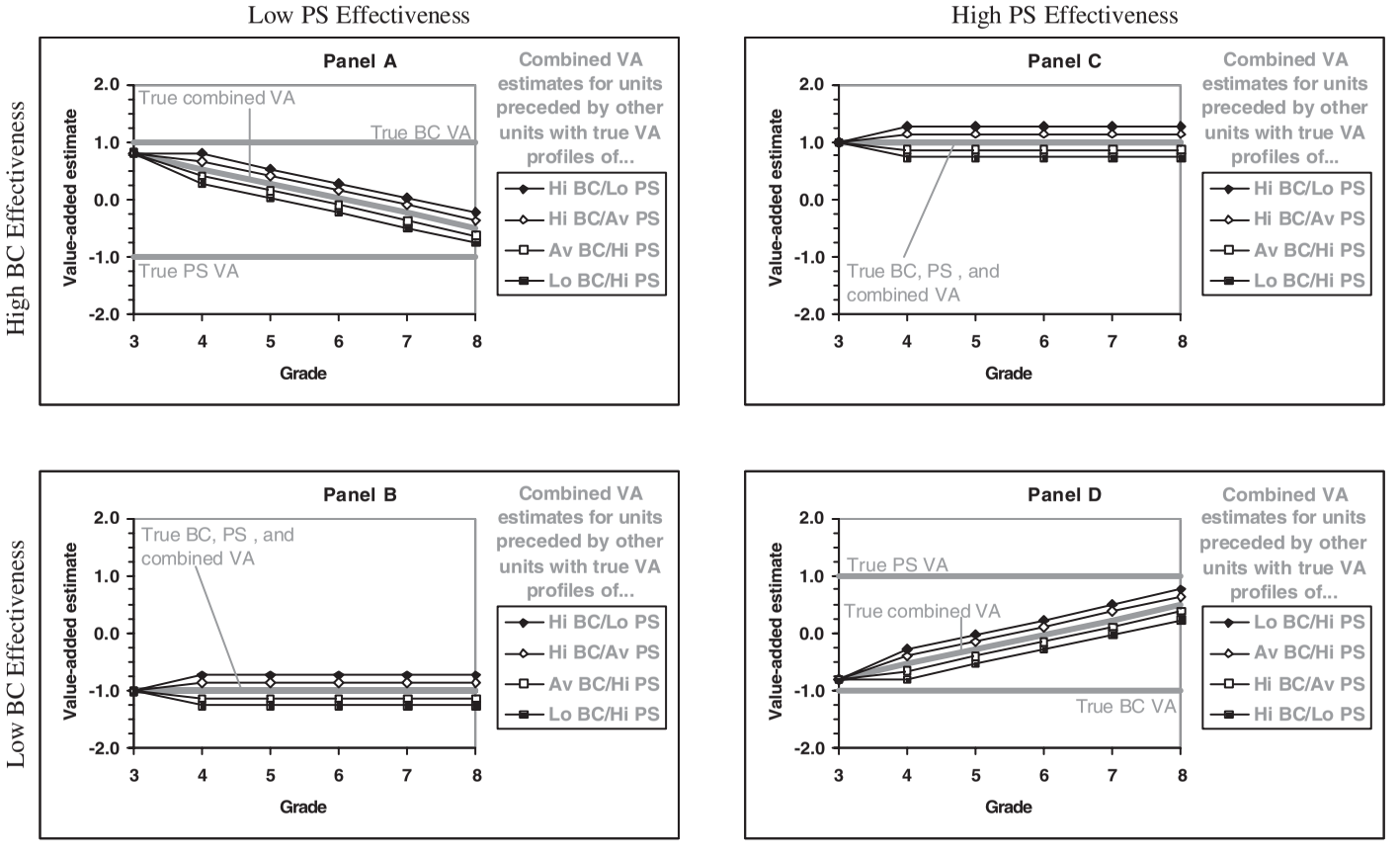

FIGURE 2. *Example effects of linear construct shift on VAA estimates.*

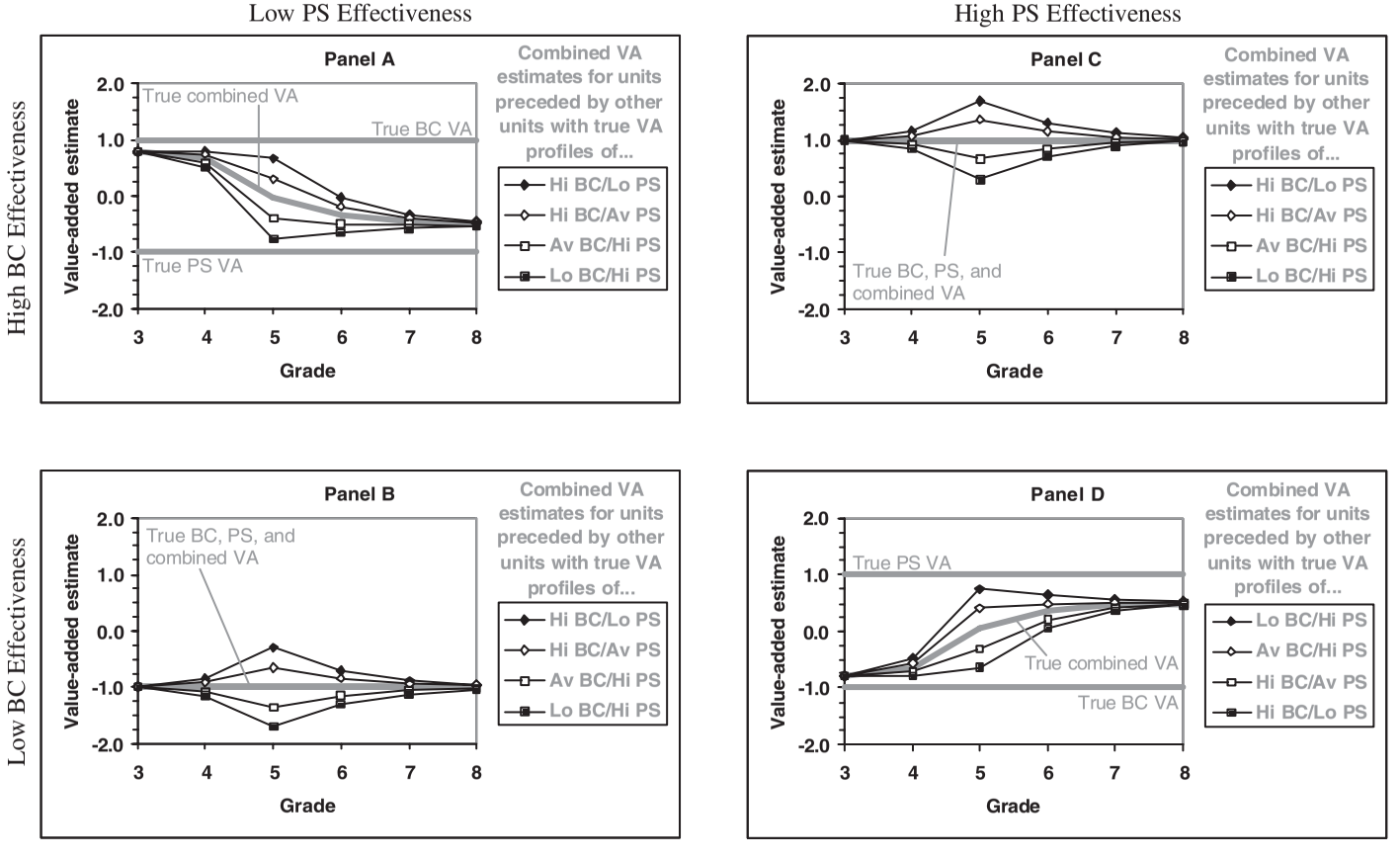

FIGURE 3. *Example effects of nonlinear construct shift on VAA estimates.*

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.

Table 2 shows the upper bounds on reliability of empirically multidimensional value-added estimates in the simplest case scenario of only two constructs combining to create a single score. The reliabilities presented in Table 2 are calculated using Equation 42. It should be noted that the reliabilities in Table 2 are upper bounds on the actual reliability of value-added estimates in three ways. First, they are reliabilities of asymptotic estimates of value added. Second and third, two of the assumptions used to facilitate the mathematics reduce the denominator of the reliability.

For high-stakes use of test data, acceptable reliabilities tend to be very high. For this study, "very high": is assumed to be 0.90 or higher, meaning that about 90% of the variation in estimated scores can be accounted for by variation in true scores. Because the values in Table 2 are upper bounds in three ways, the level of reliability deemed acceptable is placed higher at 0.95. All values of 0.95 or less are shaded gray, with lower values being shaded darker grays.

Even for low-stakes research purposes, very low reliabilities tend to be useless. For this study, "very low" is assumed to be 0.60 or less, and all values of 0.60 or less are inverted (printed in white lettering on dark background).

The pattern of reliabilities displayed in Table 2 shows that (a) later grades in the analysis tend to have lower reliabilities, (b) large changes in content representation are associated with unacceptable reliabilities, (c) more equal proportional representation of various constructs is associated with lower reliability of value-added estimates, and (d) higher interconstruct correlations of value added tend to ameliorate the effects of construct shift to some degree, with perfect correlations eliminating the effects.

The pattern in Table 2 shows that for high-stakes use, the effects of construct shift cannot be reasonably ignored except in the most fortuitous circumstances where early grade status, balanced proportional representation, small change in proportion, and/or high (near unity) interconstruct correlations combine to eliminate the effects of construct shift effectively. For high-stakes use, Table 2 also shows that in most reasonable circumstances, the upper bound on reliability is so low as to render value-added estimates not only trivially informative, but damaging for high-stakes use.

For low-stakes research purposes, empirically multidimensional VA estimates have a wider range of application. Table 2 shows that research purposes are reasonably served where any two of the following three circumstances apply: (a) shift in content is small from the prior grade, (b) there are few prior grades included in the analysis, and (c) the intraconstruct correlation of value added is high.

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

TABLE 2
Simplest-Case (Two-Construct) Upper Bounds on Reliability of Empirically Multidimensional Value-Added Estimates.

Proportional Representation[a]	Change in Representation[a]	# of Prior Grades in the Analysis	Interconstruct Value-Added Correlation										
			0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.2 (Unbalanced)	0.1 (Small)	1	0.97	0.98	0.98	0.98	0.99	0.99	0.99	0.99	1.00	1.00	1.00
		2	0.89	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	1.00
		3	0.79	0.81	0.84	0.86	0.88	0.90	0.92	0.94	0.96	0.98	1.00
		4	0.68	0.71	0.74	0.78	0.81	0.84	0.87	0.90	0.94	0.97	1.00
		5	0.58	0.61	0.65	0.69	0.73	0.77	0.81	0.86	0.90	0.95	1.00
	0.2 (Medium)	1	0.89	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	1.00
		2	0.68	0.71	0.74	0.78	0.81	0.84	0.87	0.90	0.94	0.97	1.00
		3	0.49	0.52	0.56	0.61	0.65	0.70	0.75	0.81	0.87	0.93	1.00
		4	0.35	0.38	0.42	0.46	0.51	0.57	0.63	0.70	0.79	0.88	1.00
		5	0.25	0.28	0.32	0.36	0.40	0.46	0.52	0.60	0.70	0.83	1.00
	0.3 (Large)	1	0.79	0.81	0.84	0.86	0.88	0.90	0.92	0.94	0.96	0.98	1.00
		2	0.49	0.52	0.56	0.61	0.65	0.70	0.75	0.81	0.87	0.93	1.00
		3	0.30	0.33	0.36	0.41	0.45	0.51	0.57	0.65	0.74	0.86	1.00
		4	0.19	0.22	0.24	0.28	0.32	0.37	0.43	0.51	0.62	0.77	1.00
		5	0.13	0.15	0.17	0.20	0.23	0.27	0.33	0.40	0.51	0.68	1.00
0.5 (Balanced)	0.1 (Small)	1	0.96	0.97	0.97	0.98	0.98	0.99	0.99	0.99	1.00	1.00	1.00
		2	0.86	0.88	0.90	0.92	0.94	0.95	0.96	0.97	0.98	0.99	1.00
		3	0.74	0.77	0.81	0.84	0.87	0.89	0.92	0.94	0.96	0.98	1.00
		4	0.61	0.66	0.70	0.74	0.78	0.82	0.86	0.90	0.93	0.97	1.00
		5	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
	0.2 (Medium)	1	0.86	0.88	0.90	0.92	0.94	0.95	0.96	0.97	0.98	0.99	1.00
		2	0.61	0.66	0.70	0.74	0.78	0.82	0.86	0.90	0.93	0.97	1.00
		3	0.41	0.46	0.51	0.56	0.62	0.68	0.74	0.80	0.86	0.93	1.00
		4	0.28	0.32	0.37	0.42	0.48	0.54	0.61	0.69	0.78	0.88	1.00
		5	0.20	0.23	0.27	0.32	0.37	0.43	0.50	0.59	0.69	0.83	1.00
	0.3 (Large)	1	0.74	0.77	0.81	0.84	0.87	0.89	0.92	0.94	0.96	0.98	1.00
		2	0.41	0.46	0.51	0.56	0.62	0.68	0.74	0.80	0.86	0.93	1.00
		3	0.24	0.27	0.32	0.36	0.42	0.48	0.55	0.64	0.74	0.85	1.00
		4	0.15	0.18	0.21	0.24	0.29	0.34	0.41	0.50	0.61	0.77	1.00
		5	0.10	0.12	0.14	0.17	0.21	0.25	0.31	0.39	0.50	0.68	1.00

[a]Of either construct

pure construct; therefore, VAA models are unlikely to be of practical utility for supporting an interpretation of value added to a single construct.

Empirically unidimensional score scales support a *static construct mix* interpretation of value added to student gains. It is possible to create an empirically unidimensional score scale by carefully constructing parallel forms. VAA models built upon empirically unidimensional 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 con-

tent 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 are 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 used, 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. Additional contributors to the impact of the distortions include the number of prior grades included in the analysis, interconstruct correlations of value added, and a relatively balanced mix of constructs on the test.

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 fortuitous circumstances where strong correlations are also combined with small changes in content from the previous grade, small numbers of prior grades in the analysis, and/or relatively unbalanced construct mix on the test.

For high-stakes use, estimates of value added derived from vertically scaled achievement data are not only uninformative, but can be damaging: the upper bound on reliability of the estimates is simply unacceptable for high-stakes use.

For low-stakes research purposes, estimates of value added derived from vertically scaled achievement data can have some valid application where any two of the following three conditions apply: (a) shift in content is small from the prior grade, (b) there are few grades included in the analysis (e.g., two or three), and (c) the intraconstruct correlation of value added is high. Conditions (a) and (c) are empirical questions that should be researched before applying VA methodology for low-stakes research purposes to ensure that the analyses will be informative.

With current technology, there are no vertical score scales that can be validly used in high-stakes analyses for estimating 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 high-stakes 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." Even this approach does not address the problems of using vertically scaled achievement data to estimate value

added by units instructing students who are far behind or far ahead of the range measured well by the appropriate grade level test that are identified in this study.

This article strengthens Yen's (1986) caution against comparisons of scores from widely separated test levels to caution against high-stakes comparisons of *adjacent* grade level scores if a significant shift in content mix occurs across the grade levels of the test.

Implications, Limitations, and Future Research

A serious (but reasonable) implication of this study is to all but eliminate the high-stakes use of value-added accountability systems based on vertically scaled student achievement data. The only high-stakes use likely to address the issues raised in this study is to measure student achievement on parallel forms at the beginning (or end) of adjacent grades, analyzing difference scores from the parallel forms. Until VAA technology addresses the problems of construct shift identified in this study, this is the only defensible use of vertically scaled achievement data for high-stakes VAA use. Even this use does not address the problems identified in this study for units instructing students far behind or far ahead of the range of achievement measured well by the appropriate grade level test. These units are expected to facilitate growth outside this range of achievement. Therefore, this approach should be applied only with great caution, taking care to avoid the misapplication of this approach to units expected to facilitate growth that is far ahead or far behind the norm for their grade level.

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 nonlinear 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, 1989a, 1989b, 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.

An additional approach to reduce construct shift over vertical scales that merits study is embedding a large majority of upper- and lower-grade items at each grade level (using matrix sampling across a large number of forms), combined with the creation of a separate vertical scale for each pair of adjacent grades. This approach may minimize construct shift across adjacent pairs of grades, while allowing for shift between the scales developed for adjacent pairs of grades.

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 article 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.

Notes

[1]This differs from the traditional definition of essential unidimensionality (see Stout, 2002) in two ways: (a) 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 (b) the definition involves a longitudinal component.

[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.

[3]This provides an upper bound on reliability of one-time samples of estimates of value added.

[4]This assumption provides an additional upper bound on the reliability, since prior units (or units generally in the same locale) are more likely to have similar effectiveness profiles, adding to the variance of the distortion which is only in the denominator of the reliability.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association (APA).

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association (AERA).

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002, Winter). Standards for educational accountability systems. *The CRESST Line,* 1–4.

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

Barnard, J. J. (1996, September). *In search for equity in educational measurement: Traditional versus modern equating methods.* Paper presented at the National Conference of the Association for the Study of Evaluation in Education in South Africa, Pretoria, South Africa.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17*(1), 5–9.

Bryk, A. S., Thum, Y. M., Easton, J. Q., & Luppescu, S. (1998). Assessing school academic productivity: The case of Chicago school reform. *Social Psychology of Education, 2,* 103–142.

Camilli, G. (1999). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. *Journal of Educational Measurement, 36*(1), 73–78.

Cliff, N. (1991). Ordinal methods in the assessment of change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 34–46). Washington, DC: American Psychological Association.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory.* New York: Holt, Rinehart, & Winston.

Fuhrman, S. H., & Elmore, R. F. (2004). *Redesigning accountability systems for Education.* New York: Teachers College Press.

Goertz, M. E., Duffy, M. C., & Le Floch, K. C. (2001). *Assessment and accountability systems in the 50 states: 1999–2000* (No. RR-046). Philadelphia, PA: Consortium for Policy Research in Education.

Herman, J. L., Brown, R. S., & Baker, E. L. (2000). *Student assessment and performance in the California public school system.* Los Angeles: Center for the Study of Evaluation (CSE), National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001). *Toward a framework for validating gains under high stakes conditions* (Technical Report 551). Los Angeles: Center for the Study of Evaluation, UCLA.

Lewis, A. (2001). *2000 CRESST Conference proceedings: Educational accountability in the 21st century* (No. CSE Technical Report 549). Los Angeles: Center for the Study of Evaluation (CSE), National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems* (No. CSE Technical Report 539). Los Angeles: Center for the Study of Evaluation (CSE), National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20,* 389–404.

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

Michigan State University Education Policy Center. (2002). *State accountability frameworks.* Retrieved August 19, 2004, from http://www.epc.msu.edu/publications/accountability/StateAccountability.pdf

Millman, J. (Ed.). (1997). *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* Thousand Oaks, CA: Corwin Press.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and practice.* Princeton, NJ: Educational Testing Service Policy Information Center.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425, (2002).

Olson, L. (2002, November 20). Education scholars finding new "value" in student test data. *Education Week,* 1–14.

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

Reckase, M. D. (1989a, August). *Controlling the psychometric snake: Or, how I learned to love multidimensionality.* Paper presented at the Annual Meeting of the American Psychological Association, New Orleans, LA.

Reckase, M. D. (1989b, March). *Similarity of the multidimensional space defined by parallel forms of a mathematics test.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Reckase, M. D. (1998, October). *Investigating assessment instrument parallelism in a high dimensional latent space.* Paper presented at the Annual Meeting of the Society of Multivariate Experimental Psychology, Woodcliff Lake, NJ.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*(3), 193–203.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*(4), 361–373.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*(3), 299–311.

Sanders, W. L., Saxon, 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 evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Schacter, J. (2001). *Teacher performance-based accountability: Why, what and how.* Santa Monica: Milken Family Foundation.

Schulz, E. M., & Nicewander, W. A. (1997). Grade equivalent and IRT representations of growth. *Journal of Educational Measurement, 34*(4), 315–331.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331–354.

Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53–57). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Stout, W. (2002). Psychometrics: From practice to theory and back. 15 Years of nonparametric multidimensional IRT, DIF/Test equity, and skills diagnostic assessment. *Psychometrika, 67*(4), 485–518.

Thum, Y. M. (2002). *Measuring progress towards a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis.* Santa Monica: Milken Family Foundation.

Thum, Y. M. (2003). *No Child Left Behind: Methodological challenges & recommendations for measuring adequate yearly progress* (No. 590). Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles (CSE).

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics, 24,* 398–412.

Webster, W. J., Mendro, R. L., & Almaguer, T. D. (1994). Effectiveness indices: A "value-added" approach to measuring school effect. *Studies in Educational Evaluation, 20,* 113–45.

Westat & Policy Studies Associates. (2001). *The longitudinal evaluation of school change and performance (LESCP) in Title I schools, Vol. 1: Executive summary.* Washington,

DC: U.S. Department of Education, Office of the Deputy Secretary, Planning and Evaluation Service.

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

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*(4), 399–325.

Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the national assessment of educational progress. *Journal of Educational Statistics, 17*(2), 205–218.

Author

JOSEPH A. MARTINEAU is Psychometrician, Office of Educational Assessment and Accountability, Michigan Department of Education, P.O. Box 30008, Lansing, MI 48909; martineauj@michigan.gov. His areas of interest and specialty are large-scale assessment, value added, growth models, and psychometrics.

Manuscript received June 3, 2004
Revision received November 1, 2004
Accepted December 28, 2004