An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance

Carmen D. Tekwe, Randy L. Carter, Chang-Xing Ma, James Algina, Maurice E. Lucas, Jeffrey Roth, Mario Ariet, Thomas Fisher and Michael B. Resnick

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2004 29: 11
DOI: 10.3102/10769986029001011

Journal of Educational and Behavioral Statistics
Spring 2004, Vol. 29, No. 1, pp. 11–36

An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance

Carmen D. Tekwe
Johns Hopkins University

Randy L. Carter
University at Buffalo

Chang-Xing Ma
James Algina
University of Florida

Maurice E. Lucas
Alachua County School Board

Jeffrey Roth
Mario Ariet
University of Florida

Thomas Fisher
Fisher Education Consulting, Inc

Michael B. Resnick
University of Florida

Hierarchical Linear Models (HLM) have been used extensively for value-added analysis, adjusting for important student and school-level covariates such as socioeconomic status. A recently proposed alternative, the Layered Mixed Effects Model (LMEM) also analyzes learning gains, but ignores sociodemographic factors. Other features of LMEM, such as its ability to apportion credit for learning gains among multiple schools and its utilization of incomplete observations, make it appealing. A third model that is appealing due to its simplicity is the Simple Fixed Effects Models (SFEM). Statistical and computing specifications are given for each of these models. The models were fitted to obtain value-added measures of school performance by grade and subject area, using a common data set with two years of test scores. We investigate the practical impact of

The authors wish to thank Richard Tate, Florida State University and other members of the Department of Education's *Ad Hoc* Committee on Value-Added Assessment for their guidance and stimulating discussions concerning this research. We thank Dr. Howard Wainer, Editor, and three anonymous referees for their comments on a previous version of this article. This research was partially funded by a contract (Project # 011-90950-00004) with the Florida Department of Education.

differences among these models by comparing their value-added measures. The value-added measures obtained from the SFEM were highly correlated with those from the LMEM. Thus, due to its simplicity, the SFEM is recommended over LMEM. Results of comparisons of SFEM with HLM were equivocal. Inclusion of student level variables such as minority status and poverty leads to results that differ from those of the SFEM. The question of whether to adjust for such variables is, perhaps, the most important issue faced when developing a school accountability system. Either inclusion or exclusion of them is likely to lead to a biased system. Which bias is most tolerable may depend on whether the system is to be a high-stakes one.

Keywords: *empirical comparison, hierarchical linear model, layered mixed effects model, simple fixed effects model, simple versus complex methods, value-added assessment*

1. Introduction

Several methods of assessing school performance based on standardized test scores have been proposed over the last 35 years. The earliest of these methods utilized only students' scores from the current year (i.e., status-score) to estimate school effects on student performance (Coleman, Campbell, & Kilgore, 1982). Status-based methods typically rely on regression models, which include school effects that are assumed fixed. These methods may or may not include student or school variables that influence test scores. The distinguishing characteristic of status-based methods is the absence of adjustment for students' incoming knowledge level. Specifically, previous year's score is not controlled when estimating school effects. The obvious deficiency of such methods is that differences among schools in average knowledge of incoming students would confound the assessment of instructional quality at each school. This aspect of the status-score methods is especially undesirable when assessing the quality of instruction by grade level, because while the school might be responsible for a students incoming math score in the third grade, for example, the 3rd-grade math teachers are not.

Because of this flaw in status-based methods, alternatives that adjust for incoming differences in knowledge level or ability are generally preferred. Aitkin and Longford, for example, stated that "the minimal requirement for valid institutional comparison is an analysis based on individual level data which adjusts for intake differences" (Aitkin, & Longford, 1986). Sanders suggested that a statistical method for measuring the influence of districts, schools and teachers on student learning that focuses on student improvement rather than absolute scores is the "only fair, reasonable thing to do if you are going to have an accountability system" (Olson, 1998). Methods that adjust for incoming knowledge level produce value-added assessments of school performance. ("Value-Added" is a term used to label methods of assessment of school/teacher performance that measure the knowledge gained by individual students from one year to the next and then use that measure as the basis for a performance assessment system. It can be used more gen-

erally to refer to any method of assessment that adjusts for a valid measure of incoming knowledge or ability).

A popular approach to value-added assessment has relied on Hierarchical Linear Models (HLM) analysis (Aitkin & Longford 1986; Goldstein 1997; Phillips & Adcock 1996, Raudenbush & Bryk, 1986). The hierarchical linear models that have been studied in the literature (i.e., Hierarchical Linear Mixed Models: HLMM) are special cases of the general mixed effects model (Littell et al., 1996) and are distinguished from corresponding fixed effects models (FEM) by the fact that school effects are assumed to be random. That is, in HLMM, schools are assumed to be a random sample from a larger population of schools (perhaps conceptual), whereas in fixed effects models they are taken to be the fixed population of schools to be graded. A value-added assessment of school performance can be derived from either HLMM or FEM analysis of change scores (current year score minus previous year score) or of status scores with intake score (usually last year's test score) included as a covariate.

An alternative mixed model, called the Layered Mixed Effects Model (LMEM), was suggested by Sanders and Horn (1994), to estimate school effects on student learning gains, and is the foundation of the Tennessee Value-Added Assessment System (TVAAS). The LMEM includes neither a direct measure of gain nor a measure of incoming knowledge/ability as a covariate. It, nevertheless, does produce value-added measures of school effects by utilizing the information in non-zero covariance between test scores at different times (Sanders & Horn, 1994). Carter, et al. (2001) and McCaffrey, et al. (2003) independently demonstrated that the LMEM can be viewed as a model for change scores with random school effects. An LMEM can be specified to either analyze multiple subject area test scores simultaneously (i.e., multivariate LMEM) or separately (i.e., univariate LMEM).

There is a natural desire on the part of the public and the educational establishment that implementation of school accountability systems involve simple methods understood by many, not just those with extensive methodological training. Systems based on very simple value-added measures such as school specific mean change scores minus the district-wide mean of mean change scores, which are obtained from the simple change score fixed effects model (SFEM), are to be preferred if they are "just as good as" systems that are based on more complex measures. Thus, there is a "burden of proof" on value-added measures developed from complex FEM, HLMM, or LMEM.

This article focuses on assessing whether the value-added measurements obtained from the LMEM or HLMM provide notably different results than those based on SFEM and whether the results from the LMEM and HLMM differ notably. Four models and their distinguishing features are described in detail in Sections 3 and 4 (i.e., the simple change score FEM; the simple change score HLMM with no covariates and a random intercept only; the demographic and intake adjusted change score HLMM; and the multivariate LMEM). The models were fitted and the value-added measures obtained were correlated to determine whether the distinguishing features of each model produced substantially different results. To our knowledge, this is the

first comparative study of LMEM, HLMM and FEM value-added methods based on the analysis of a common data set.

2. Sample Description

Separate analyses were carried out for each of three elementary school grade cohorts (i.e., 3rd–5th grades) in 1999 in a medium sized Florida school district with 22 elementary schools to be graded. Consecutive year math and reading scores on the Iowa Test of Basic Skills (ITBS) from 1998 and 1999 were analyzed. Students who were enrolled in special education programs (except those for gifted, speech or language impaired, or homebound students) and those who were in English for Speakers of Other Languages (ESOL) classes and had been in ESOL for less than two years were excluded from all analyses.

Our goals to assess the impact of various features of the different models and to do so in a common database were somewhat at odds. One of the advantages of the multivariate LMEM is that it allows the use of incomplete data, i.e. a student need not have subject specific test scores in both years for their data to be included in the analysis. In contrast, the analysis of change scores in other models requires that observations be available from both years. Thus, it is not possible to achieve both goals while using a single common dataset. Consequently, we fitted each of the four models twice, once using all available data for each model and a second time using only students with both math and reading test scores in both 1998 and 1999 (i.e., complete data).

Because the LMEM utilizes some observations that cannot be used when fitting the models with change score outcome, the sample sizes for the analysis of change scores were smaller than for LMEM when analyzing all available data for each model. In the analyses of complete data, a common complete data sub sample was used to fit all models.

A total of 6,707 students were available for use in the LMEM analyses after the exclusions (2,310 for the analysis of 3rd-grade test scores in 1999, 2,307 for the 4th-grade analysis, and 2,090 for fifth grade). The sample sizes available for fitting the three change score models are given in the "Change Score" columns of Table 1. Table 1 also presents sample sizes, summary statistics of test scores, the percentages of minority and of poverty students for each year by subject and study cohort. Table 2 contains 1999 mean scores, sample sizes, percentage of minority students, and percentage poverty students by school and grade in 1999, ordered by percentage poverty among third graders.

The minority status of a student was defined as Black or non-Black race. In this district, almost all students are non-Hispanic blacks or Whites. Most of the *relatively* small numbers of Hispanic students are White. Most students of other races are Asians but are relatively few in number. They were grouped in the "non-minority" category because of the similarity of their test score profiles. Poverty status was based on whether or not the student received free or reduced lunch subsidy. Table 1 also provides the overall percentages of minority and poor students for the elementary schools in the district.

TABLE 1
Sample Size, Mean ITBS and Standard Deviation by Subject, Cohort and Year, and Percent Minority and Percent Poverty in 1999 by Cohort

		Mathematics			Reading			Demographics in 1999	
Analysis Cohort		1998 score	1999 score	Change score	1998 score	1999 score	Change score	Percent Minority	Percent Poverty
A[1]	*N*	2076	1892	1669	2075	1886	1669	1904	1904
(*n* = 2310)	*M*	171.4	191.1	17.7	170.4	186.6	13.7	35.5	52.7
	SD	22.2	25.0	17.7	20.5	25.0	16.1		
B[2]	*N*	2101	1812	1622	2103	1815	1628	1818	1818
(*n* = 2307)	*M*	188.0	210.6	19.3	184.0	209.5	22.0	36.3	48.5
	SD	25.9	29.0	19.0	26.0	30.2	18.5		
C[3]	*N*	1926	1383	1229	1922	1382	1226	1386	1386
(*n* = 2090)	*M*	206.5	224.8	16.4	204.9	217.2	9.8	35.1	49.5
	SD	30.2	34.0	20.9	31.2	29.1	18.4		
Overall	*N*	6103	5087	4520	6100	5083	4523	5108	5108
(*n* = 6707)								35.7	50.3

Notes.
[1] Analysis Cohort A consists of all children in third grade in one of the 22 schools to be graded in 1999 with at least one test score, records with their matching test score in 1998, and children in the 22 schools in second grade in 1998 who were not in any of the 22 schools in 1999. The data from this cohort forms the "all available data" dataset for the LMEM analysis of third graders in 1999.
[2] This is the cohort whose data forms the "all available data" dataset for the LMEM analysis of fourth graders in 1999, where "all available data" is defined as for Cohort A above.
[3] This is the cohort whose data forms the "all available data" dataset for the LMEM analysis of fifth graders in 1999, where "all available data" is defined as for Cohort A above.

Two results in Table 1 call for explanation. First, there were fewer scores in 1999 than in 1998 in all cases. Second, the difference between the average scores in 1999 and 1998 (based on all data) was larger than the average change score. These results were probably due to the fact that students who were retained in 1998 were not in the samples for the next higher grade in 1999 and the fact that retained students scored lower on average than those who were promoted. In the next section we present the precise mathematical and computing descriptions of the four models studied and discuss their distinguishing characteristics.

3. Model Specifications

The four models studied were the SFEM (Model 1), the simple unadjusted change score HLMM (UHLMM) with random intercept (Model 2), a demographic and intake score adjusted HLMM (AHLMM) with outcome defined by the change score from 1998 to 1999 and with student-level and school-level covariates (Model 3), and Sanders' and Horn (1994) multivariate LMEM (Model 4). Precise specifications of these models are presented in the next four subsections. In practice, computing and

TABLE 2

Mean Scores, Percent Minority and Poverty, and Number of Students by Grade within Schools in 1999

	Third Grade					Fourth Grade					Fifth Grade				
School	Math	Reading	Percent Minority	Percent Poverty	*N*	Math	Reading	Percent Minority	Percent Poverty	*N*	Math	Reading	Percent Minority	Percent Poverty	*N*
1	166.4	165.0	79.2	91.7	48	181.1	177.0	78.9	89.5	38	197.1	186.6	81.0	92.9	42
2	159.6	157.2	73.8	90.2	61	181.0	173.8	75.9	79.6	54	194.9	200.1	83.3	88.1	42
3	159.1	164.4	75.4	86.0	57	180.9	175.5	64.1	71.9	64	192.9	194.5	56.0	80.0	50
4	155.5	162.4	87.4	83.9	87	169.9	166.9	94.4	91.7	72	193.3	189.9	92.6	75.9	54
5	164.3	162.5	37.3	80.4	51	183.6	178.7	38.6	61.4	57	197.7	199.6	21.7	67.4	46
6	169.8	164.9	76.5	76.5	68	178.6	170.3	67.9	83.9	56	193.2	193.6	70.4	76.1	71
7	155.7	162.0	68.0	76.0	75	182.7	178.8	65.8	63.3	79	198.0	200.9	64.1	67.9	78
8	165.2	165.0	53.7	75.8	95	186.1	180.9	48.0	64.7	102	205.2	203.5	45.5	61.0	77
9	175.4	173.7	31.1	75.6	45	187.2	187.3	33.3	62.7	51	210.2	223.3	34.7	73.5	49
10	178.1	171.0	13.9	75.0	36	194.5	188.9	11.1	77.8	36	204.8	199.0	29.4	55.9	34
11	167.1	169.4	36.7	74.7	79	180.3	181.7	47.4	70.5	78	205.7	202.8	42.3	71.2	52
12	177.0	172.9	26.5	63.2	68	187.6	186.3	19.4	59.7	72	201.2	207.8	15.8	51.3	76
13	174.2	172.7	28.3	52.9	191	194.0	189.8	21.6	46.2	171	205.2	203.3	19.8	41.2	131
14	175.6	174.9	23.7	48.5	97	193.1	189.4	28.8	36.9	111	212.7	211.4	26.7	41.6	101
15	170.8	174.9	14.5	39.1	110	195.5	188.0	20.2	38.3	94	—	—	—	—	—
16	175.1	170.1	25.6	38.4	86	191.3	186.6	39.7	47.4	78	209.6	206.5	22.4	37.3	67
17	182.8	181.4	22.9	34.3	70	200.1	199.7	23.9	23.9	67	223.5	217.7	14.3	30.2	63
18	180.3	180.6	15.8	30.3	165	196.5	193.5	22.4	32.8	116	222.8	218.0	16.8	24.8	137
19	178.8	178.0	14.6	30.3	89	203.5	204.7	16.0	11.7	94	—	—	—	—	—
20	181.4	175.9	28.6	29.6	98	199.6	195.9	31.1	33.3	90	228.1	222.4	20.6	23.5	102
21	182.8	181.6	21.4	26.5	98	203.3	194.9	23.3	25.9	116	221.0	221.0	10.5	13.2	114
22	186.1	183.8	12.3	13.8	130	206.9	202.5	13.1	14.8	122	—	—	—	—	—

statistical specifications of models are necessarily coupled. Therefore, because we used SAS for computations and because the SAS specification of at least one of the models considered was not transparent (*i.e.,* LMEM), we present both statistical and SAS specifications of each model.

3.1 Model 1 (SFEM)

The parameterization used for the SFEM was:

$$d_{ijs} = \beta_{0s} + \sum_{k=1}^{21} \beta_{1ks} S_{kij2} + \varepsilon_{ijs}, \tag{1}$$

where

$d_{ijs} = y_{ijs2} - y_{ijs1}$,
y_{ijst} = the test score on the s^{th} subject at time t for the j^{th} student, who at test date in 1999 ($t = 2$) was in the i^{th} school, $s = 1, 2$, $t = 1, 2$, $I = 1, 2, \ldots, 22$, $j = 1, 2, \ldots, n_i$,
S_{kij2} = 1, 0, or −1 as $I = k$ and $I \neq 22$, $I \neq k$ and $I \neq 22$, or $I = 22$, $k = 1, 2, \ldots, 21$, respectively,
and $\varepsilon_{ijs} \overset{\text{iid}}{\sim} N(0, \sigma^2_{\varepsilon s})$ for each given $s = 1, 2$.

The β_{1ks} coefficient in Equation 1 is interpreted as the value-added in the s^{th} subject area by the k^{th} school, which was measured by the estimate resulting from the model fit. These value-added measures can be easily calculated and understood as the difference between the school specific sample average change and the average of these average changes.

PROC GLM of SAS was used to fit Model 1 for each grade by subject combination. The SAS specification was

```
PROC GLM;
MODEL CHANGE = S1 − S21/SOLUTION;
```

where S1, S2, . . . , S21 are the 1, 0, −1,coded variables defined above.

3.2 Model 2 (UHLMM)

The simple unadjusted HLMM is the two-level HLMM defined by the following student level and school level models:

Student-level model

$$d_{ijs} = \beta_{0is} + \varepsilon_{ijs},$$

where d_{ijs} is the change score defined as in Equation 1, β_{0is} is a random intercept associated with the i^{th} school and ε_{ijs} is a random error.

School-level model

$$\beta_{0is} = \gamma_{0s} + \xi_{is},$$

where γ_{0s} the mean of the random intercepts, β_{0is}, and ξ_{is} is the random effect of the ith school on the random intercept for the s^{th} subject area. It was assumed that the ε_{ijs} and ξ_{is} were independent and that

$$\varepsilon_{ijs} \overset{\text{iid}}{\sim} N(0, \sigma^2_{\varepsilon s}),$$
$$\xi_{is} \overset{\text{iid}}{\sim} N(0, \sigma^2_{\xi s}),$$

for each fixed value of s.

To specify two-level hierarchical linear mixed models in SAS, it generally is helpful to write them in the form of the general linear mixed model. This is accomplished by substituting the right hand side of the second level model for the random coefficients in first level and produces a single equation form for the HLMM. For the UHLMM defined in this subsection, we have the following:

Single equation form

$$d_{ijs} = \beta_{0s} + \xi_{is} + \varepsilon_{ijs}. \tag{2}$$

The SAS statements used to specify this model were:

```
PROC MIXED;
CLASS STUDENT;
MODEL CHANGE =;
RANDOM INTERCEPT / TYPE = UN SUB = SCHOOL SOLUTION;
REPEATED/TYPE = UN SUB = STUDENT;
```

Value-added measures were calculated as estimates of best linear unbiased predictors (BLUPs) of the ξ_{is}, $i = 1, 2, \ldots, 22$, random school effects in each model, $s = 1, 2$, for each grade level. These values-added measures were obtained by including the SOLUTION option in the RANDOM statement above and are shrunken versions of the estimates of school effects in Model 1 above, defined in Equation 1. Readers are referred to Littell, et al. (1996), for a thorough presentation on BLUPs and calculations by SAS.

3.3 Model 3 (AHLMM)

The third model considered was an HLMM adjusted for student- and school-level covariates. The covariates included in separate analyses for each grade by subject combinations were determined from a preliminary model fitting described in the Appendix. The resulting model specification for third grade math, for example, was the following two-level HLMM:

Student level model

$$d_{ijs} = \beta_{0is} + \beta_{1s} y_{ijs1} + \beta_{2s}\text{Min}_{ij} + \beta_{3s}\text{Pov}_{ij} + \varepsilon_{ijs},$$

where $d_{ijs} = y_{ijs2} - y_{ijs}$, β_{0is} is a random intercept associated with the i^{th} school and subject area s, Min_{ij} = an indicator of minority status (Yes, No) for the j^{th} student in the i^{th} school in 1999, Pov_{ij} = an indicator of poverty (Yes, No) for the j^{th} student in the i^{th} school, β_{1s}, β_{1s}, and β_{3s}, are the fixed effects of intake score, minority status, and poverty on learning gain in subject area s, and ε_{ijs} is a random error.

School-level model

$$\beta_{0is} = \gamma_{0s} + \gamma_{1s} z_{1i} + \gamma_{2s} z_{2i} + \xi_{is},$$

where z_{1i} is the mean input score for the i^{th} school, z_{2i} is the percentage of poverty students in the i^{th} school, ζ_{is} is the random error associated with the value of the random intercept for the s^{th} subject area test and the i^{th} school in the student level model and the γ's are fixed coefficient parameters.

The assumptions concerning the within and between school error terms in this model were that the ε_{ijs} and ξ_{is} are independent, and that

$$\varepsilon_{ijs} \overset{\text{iid}}{\sim} N(0, \sigma^2_{\varepsilon s}),$$
$$\xi_{is} \overset{\text{iid}}{\sim} N(0, \sigma^2_{\xi s}),$$

for each fixed value of s.

Single equation form

The student-level and school-level models can be written in single equation form as

$$d_{ijs} = \gamma_{0s} + \gamma_{1s} z_{1i} + \gamma_{2s} z_{2i} + \beta_{1s} y_{ijs1} + \beta_{2s} \text{Min}_{ij} + \beta_{3s} \text{Pov}_{ij} + \xi_{is} + \varepsilon_{ijs}. \quad (3)$$

The following statements achieve the SAS specification of the single equation form (Equation 3) of Model 3:

```
PROC MIXED;
CLASS STUDENT MIN POV;
MODEL CHANGE = Z1 Z2 Y1 MIN POV;
RANDOM INTERCEPT/TYPE = UN SUB = SCHOOL SOLUTION;
REPEATED/TYPE = UN SUB = STUDENT;
```

The value-added measures from Model 3 were calculated as the estimated BLUP estimates of ξ_{is}, $i = 1, 2, \ldots, 22$, in each model, $s = 1, 2$, for each grade level. These estimates were calculated by SAS as a result of including the SOLUTION option in the RANDOM statement above.

3.4 Model 4 (LMEM)

The simplest form of the model used in the LMEM analysis for this application was

$$y_{ijst} = \mu_{st} + \sum_{l=1}^{t} \sum_{k=1}^{22} P_{ijkl} u_{ksl} + \varepsilon_{ijst},$$

where y_{ijst} is defined as in Model 1, Equation 1, and

μ_{st} = the population mean parameter for the s^{th} subject test score at time t,
P_{ijkl} = the proportion of academic year time spent by the j^{th} student, who was in the i^{th} school at time 2 test, in the k^{th} school during the year prior to the test at time l, $l = 1, t$,
u_{ksl} = the random effect of the k^{th} school on subject s test scores at time l,
ε_{ijst} = random within school error for the j^{th} student in the i^{th} school for the s^{th} subject at time t.

An alternative specification that put this model in the form of a general linear mixed model to be fitted by SAS PROC MIXED was

$$y_{ijst} = \sum_{s=1}^{2}\sum_{t=1}^{2} X_{ijst}\mu_{st} + \sum_{s=1}^{2}\sum_{l=1}^{t}\sum_{k=1}^{22} Z_{ijksl}u_{ksl} + \varepsilon_{ijst}, \quad (4)$$

where y_{ijst}, μ_{st}, P_{ijkl}, u_{ksl}, and ε_{ijst} are defined above, and

$$X_{ijst} = \begin{cases} 1, & \text{if the score is for the } s^{th} \text{ subject at time } t, \\ 0, & \text{if not,} \end{cases}$$

$$Z_{ijksl} = \begin{cases} P_{ijkl}, & \text{if the score is for subject } s \text{ at time } l, 1 \le l \le t, \\ 0, & \text{otherwise.} \end{cases}$$

Following the TVAAS assumptions, we assumed that the u_{ksi} and ε_{ijst} random effects were independent normally distributed with mean zero, $\text{Var}(u_{ksl}) = \sigma^2_{sl}$, $\text{Cov}(u_{ksl}, u_{k's'l'}) = 0$ for all $k \neq k'$, $s \neq s'$, or $l \neq l'$ (Note, however, that a more natural assumption would allow for correlation between subjects or times.), and $\text{Cov}(\varepsilon_{ijst}, \varepsilon_{i'j's't'}) = 0$ for all $(i, j) \neq (i', j')$ but that the covariance matrix of the ε_{ijst} otherwise was unstructured. Note that the last of these assumptions allows the LMEM to utilize information contained in intra-student correlations among test scores. This feature of the LMEM is indicated herein by the term "multivariate analysis."

Inclusion of P_{ijkl} in the definition of Z_{ijksl} in the LMEM, Equation 4 results in a partitioning of the total effect of schools attended during the year prior to the test in year t in proportion to the proportion of time spent in each school. This can be illustrated by considering two cases: (a) if the j^{th} student in the k^{th} school at the time 2 had attended only the k^{th} school since the test at time 1 and had attended only the k^{th} school for the year prior to time 1, then the second term in Equation 4 for a math score at time 2 would be $u_{k11} + u_{k12}$. For a math score at time 1 this term would be u_{k11}. Thus, the effect of the k^{th} school on the change score in this case would be the difference, u_{k12}; and (b) If, however, the student attended school k' for the first half of the year prior to the test in year 2, the second half in school k and the entire year before the test in year 1 in school k', then the second term in Equation 4 for a math score at time 2 would be $u_{k'11} + 0.5u_{k'12} + 0.5u_{k12}$. The second term would be $u_{k'11}$ for the math score at time 1. Hence, the difference, $0.5u_{k'12} +$

$0.5u_{k12}$, would measure the total effect of schools k and k' on the change score. Thus, the use of LMEM produces value-added measures (*i.e.,* estimates of u_{ksl}) that can be viewed as estimated school effects on change scores. We see in the latter case how school effects are apportioned according to the proportion of time spent in each school attended during the year prior to the test. This feature makes the LMEM an attractive choice for value-added assessment, provided that (a) a school's contribution to a student's gain in knowledge in one year is approximately proportional the proportion of time enrolled in that school during the year; and (b) attributing proportional credit for gains has a notable impact on results compared to attributing the entire gain to the school of attendance at the time of the test. It is debatable whether (a) holds, however, as students may be inclined to change schools when they are not learning much. Even when (a) is a reasonable assumption, it would not be worth the effort to apportion credit if doing so has no effect on value added assessment of school performance compared with that based on the SFEM.

The following statements specify Model 4 in SAS:

```
PROC MIXED;
CLASS STUDENT;
MODEL Y = X11 X21 X12 X22/NOINT SOLUTION;
```

where X11, X21, X12, and X22 are the X_{ijst} variables in Model 4 for different values of s and t,

```
RANDOM ZM1_1-ZM1_22/TYPE = TOEP(1) SOLUTION;
RANDOM ZR1_1-ZR1_22/TYPE = TOEP(1) SOLUTION;
RANDOM ZM2_1-ZM2_22/TYPE = TOEP(1) SOLUTION;
RANDOM ZR2_1-ZR2_22/TYPE = TOEP(1) SOLUTION;
```

where the Z variables in the SAS statements are defined by Z_{jiksl} in Model 4, and

```
REPEATED/TYPE = UN SUB = STUDENT;
```

The TOEP(1) option in the RANDOM statements specifies Toeplitz covariance structure (i.e., banded structure), with one band, for the corresponding $\boldsymbol{u}$ vector. This is the structure assumed by TVAAS. To relax the TVAAS structure and allow non-zero correlation among the elements of $\boldsymbol{u}$ for a given school (i.e., the more natural assumption), one would simply use the following RANDOM statements:

```
RANDOM ZM1_1 ZR1_1 ZM2_1 ZR2_1/TYPE = UN SOLUTION;
RANDOM ZM1_2 ZR1_2 ZM2_2 ZR2_2/TYPE = UN SOLUTION;
                        ⋮
RANDOM ZM1_22 ZR1_22 ZM2_22 ZR2_22/TYPE = UN SOLUTION;
```

in place of those above.

The value-added measures obtained from Model 4 were the estimated best linear unbiased predictors (BLUPs) of the random school effects associated with t = 2, that is the estimates of the u_{ks2} in Model 4. The estimated BLUPs of school effects

were obtained from SAS by including the SOLUTION option in the RANDOM statements for Model 4 above.

4. Comparison of Model Features

A summary of the distinguishing features of the four models studied[1] is presented in Table 3. The virtue of the Model 1 is its simplicity. It measures each school's effect as the average change score for that school minus the average of average change scores across schools. This quantity is easily calculated and understandable to all. It would be preferred over alternative value-added measures unless proven to be inferior in practice. It is known, theoretically, that the SFEM (i.e., Model 1) has potential shortcomings due to its simplicity. It is unknown, however, whether the theoretical deficiencies translate to notably inferior performance in practice. In this section we list the potential shortcomings of SFEM and discuss how each alternative model addresses some, but not all, of them.

The first potential shortcoming of the SFEM lies in a perceived theoretical deficiency. That is, SFEM does not produce shrunken estimates of value-added.[2] Second, the SFEM ignores confounding factors, such as minority status, poverty and intake score, that might unfairly bias comparisons among schools. Thirdly, it fails to apportion credit among multiple schools attended when assessing value-added. Finally, SFEM-based procedures for value-added assessment analyze subject area test scores separately and, therefore, fail to utilize the information contained in intra-student correlations among these scores.

The UHLMM (Model 2) is a mixed effects model and, hence, produced shrunken estimates of school effects. Otherwise, it is subject to the same potential deficiencies as Model 1. Model 2 was included in the comparisons to isolate the effect of shrinkage. The UHLMM was estimated using an iterative restricted maximum likelihood (REML) based, estimated generalized least squares (EGLS) procedure and produced estimated BLUPs of random school effects. The observed estimated BLUPs were shrunken estimates of school effects and were the value added measures used in comparisons.

The AHLMM (Model 3) is an HLMM with change score as an outcome and with student and school level variables as predictors. It is a mixed model and produced shrunken estimates of school effects (*i.e.,* estimated BLUPs) through REML-based EGLS. Furthermore, because Model 3 included student and school level covariates, the resulting value added measures (estimated BLUPs) were adjusted for factors that influenced student performance. The predictor variables that were included in Model 3 for each grade by subject combination are presented in Table 4 and included minority status, poverty, and intake score. If the effects of these factors are independent of quality of instruction, then the value-added measures produced by Model 3 would be preferred to those produced by models that make no adjustments and have no other compensating advantages, such as the SFEM and UHLMM.

The AHLMM, in multilevel form, facilitates interpretations of the effects of student and school level variables, either contextual or treatment (Phillips &

TABLE 3
Models Considered and Summary of Their Distinguishing Characteristics

Model identifier	Model name	Dependent variable	Intake adjusted	School effect[1]	Student-level variables included	School-level variables included	Apportions between schools	All fractured observations used[2]	Multivariate method
Model 1	SFEM	Change Score	No	Fixed	No	No	No	No	No
Model 2	HLMM (UHLMM)	Change Score	No	Random	No	No	No	No	No
Model 3	HLMM (AHLMM)	Change Score	Yes	Random	Yes	Yes	No	No	No
Model 4	MEM (LMEM)	Pre/Post-test Scores	No	Random	No	No	Yes	Yes	Yes

Note. Models with random school effects produced shrunken estimates of value-added while those with fixed school effects did not. LMEM uses all available test scores. The other three models use only test scores with a matching subject area score in the previous year. Some fractured observations are usable by the other three models. For example, those with complete math scores but a missing reading score would be used in the math analysis but not the reading analysis.
SFEM = Simple Fixed Effects Model.
HLMM (UHLMM) = Simple Unadjusted Change Score.
HLMM (AHLMM) = Demographic and Intake Adjusted Change Score.
MEM (LMEM) = Multivariate Layered.

TABLE 4
Model Building Results for the Demographic and Intake Adjusted Change Score HLMM (Model 3)[1]

	Elementary School					
Student Variable	Grade 2 to 3		Grade 3 to 4		Grade 4 to 5	
	Math	Reading	Math	Reading	Math	Reading
Intercept	179.84	106.51	64.84	59.34	150.46	80.89
Input score	–0.26	–0.13	–0.21	–0.19	–0.24	–0.30
Minority status	–5.56	–6.27	–6.65	–30.82	–44.48	–3.96
Poverty	–4.12	–3.26	–5.09	–4.97	–5.00	–3.76
Minority status*input				0.13	0.19	
School variable						
Mean input	–0.58	–0.35			–0.34	
% Minority		10.44				
% Poverty	–25.70	–22.07			–26.41	–9.76
% Mobility				34.86	46.78	

Note. Model coefficient estimates for the model building obtained from the results of the model strategy described in the Model Fitting section and the Appendix. A cell was left blank if the corresponding variable did not enter the corresponding model (i.e., was not significant). All the HLMs had random intercepts only.

Adcock, 1996). It should be noted that, for every HLMM, a corresponding FEM is obtained by treating the random school effects in the single equation form of the HLMM as fixed. Thus, it is clear that the SFEM can be extended to adjust for important student or school level variables in a way comparable to an HLMM. The extended SFEM then loses the virtue of simplicity, however, and also does not produce shrunken estimates of school effects. Nevertheless, there are several disadvantages of HLMM compared with the corresponding FEM. First, estimation of parameters involves a computationally intense iterative procedure that sometimes fails to produce estimates of school effects because it fails to converge (this can happen, for example, when the estimate of $G = \mathrm{Var}(u)$ is not positive definite.) Also, HLMM involve more complex statistical methods, such as estimated generalized least squares estimation of fixed effects and best linear unbiased prediction of random effects, that are not well known to many with a need to understand. Assuming convergence, an HLMM would be preferred to the corresponding FEM only if shrinkage has an impact in practice. It would be preferred to SFEM if either shrinkage or adjustment for covariates has an impact. A comparison of the UHLMM with SFEM was made to assess the impact of shrinkage, while a comparison of AHLMM was made to assess the impact of significant student and school level covariates.

The LMEM (Model 4) has several appealing features that address most of the potential shortcomings of the SFEM. Like all mixed models, it produces shrunken estimates of value-added measures. Additionally, it is the only model proposed to date that apportions credit for learning gains to multiple schools attended. Further-

more, it makes full use of the available data by using all incomplete, or fractured, observations. The LMEM also allows multivariate analysis of several subjects simultaneously, thereby accounting for intra-student correlation between math and reading scores, for example. A disadvantage of LMEM, relative to SFEM, lies in the fact that the estimation procedure is complex (REML-based EGLS and BLUPs). Furthermore the estimation procedure does not always converge. Non-convergence, for example, can be a problem when the covariance matrix of u is nearly singular.

Models 3 and 4 each have different features that eliminate different sets of potential deficiencies in the less complex SFEM (Model 1). Neither, however, is totally satisfactory, each having its unique strengths and weaknesses. The AHLMM (Model 3) allows for easy adjustment for student and school level covariates while the LMEM (Model 4) does not. Model 4, on the other hand, is multivariate, apportions school effects, and utilizes all fractured observations. It should be noted that either the AHLMM or the LMEM could be modified to address all of the potential shortcomings of the SFEM. Unfortunately, neither full-featured AHLMM nor LMEM have been developed to date to address all of the concerns. Modification of the AHLMM to handle multivariate observations and to utilize fractured observations is straightforward. It is not obvious, however, how to modify the AHLMM to apportion credit to multiple schools attended. It is also not obvious how to modify the LMEM to include student or school level covariates. Such modifications are left to future research.

5. Questions Relevant to the Choice of Models

In practice, school districts or state departments of education must struggle with many issues when developing school accountability systems. It is generally accepted among experts that value-added systems are desirable. The theoretically preferred methods, however, are quite complex and produce value-added measures that are not readily understandable. It is not surprising that there has been a reluctance to adopt them. This study was designed to provide useful information to those who must choose between competing methods. The main questions considered were whether complex methods based on either the AHLMM (Model 3) or the LMEM (Model 4) produced value-added measures that were notably different from the simple and easily understood ones produced by the SFEM (Model 1). Strong agreement of the value-added measures from either with those from Model 1 would eliminate Model 3 or Model 4, respectively, from consideration as the foundation of a value-added assessment system.

The three specific questions of primary interest were:

1. What were the collective effects of shrinkage, multivariate analysis, apportioning of credit among multiple schools attended, and the use of all fractured observations? (Model 4 vs. Model 1);
2. Was there a notable collective effect of shrinkage and inclusion of student and school level covariates in the AHLMM on value-added assessment? (Model 3 vs. Model 1); and

3. Was there disagreement in value-added results from the multivariate LMEM and the AHLMM? (Model 4 vs. Model 3).

Several additional questions of secondary interest were:

4. Did shrinkage alone have an impact on value-added assessment? (Model 2 vs. Model 1);
5. Did the use of multivariate analysis of a model that explicitly acknowledges potential intra-student correlations between subject area test scores, i.e. LMEM, have an important impact compared with an analysis that ignores such correlations? (Model 4 vs. Model 2);
6. Did the inclusion of student and school covariates in the AHLMM have an impact on value-added assessment? (Model 3 vs. Model 2).

To answer these questions, the models described in Section 3 were estimated. Value-added measures were obtained for each school under each model. Correlations between the value-added measures from the appropriate models were used to answer each of the six questions posed in this section.

6. Results

The results of the analyses of "complete" and "available" data were essentially the same. We therefore report only the results from the analysis of all available data for each model fitted. The coefficients on the significant variables that were included in Model 3 after the model-building strategy was completed are presented in Table 4 (See the Appendix for information on the model building strategy). It should be noted that only minority status, poverty and *intake* score entered these models consistently across grades and subject areas. Other variables entered sporadically, and some may have been the result of Type I statistical errors.

The results of correlating value-added measures of school effects from the models relevant to the questions in the previous section are given in Table 5. The answers to the three primary questions are:

Question 1:

The global impact of using the multivariate LMEM compared to the SFEM was small, in this study of two years of data. The correlations of the value-added measures from Model 4 with those from Model 1 ranged from 0.91 to 0.98. It is possible that greater discrepancy of results would be found in studies of three or more years of data. McCaffrey, et al. (2003), suggested that allowing for cross-time correlations in the LMEM might mitigate the effects of omitting some covariates. If so, then the results from the LMEM in a study of three or more times would match more closely those from the AHLMM and less closely those from the SFEM than in the current study. Thus, the question of whether LMEM and SFEM produce nearly interchangeable results when data from three or more times is analyzed merits further study. If only two years of data are to be used, however, the result is

TABLE 5
Table of Correlations Measuring Agreement of Model Results

Grade/Subject	SFEM vs. UHLMM Model 1 vs. Model 2	SFEM vs. AHLMM Model 1 vs. Model 3	SFEM vs. LMEM Model 1 vs. Model 4	UHLMM vs. AHLMM Model 2 vs. Model 3	UHLMM vs. LMEM Model 2 vs. Model 4	AHLMM vs. LMEM Model 3 vs. Model 4
Third Grade						
Math	1.00	0.60	0.98	0.61	0.98	0.57
Reading	1.00	0.70	0.98	0.72	0.99	0.72
Fourth Grade						
Math	0.98	0.96	0.96	0.95	0.96	0.94
Reading	0.97	0.83	0.91	0.87	0.94	0.71
Fifth Grade						
Math	0.99	0.73	0.96	0.71	0.97	0.65
Reading	0.98	0.72	0.98	0.77	0.97	0.74

Note.
SFEM = Simple Fixed Effects Model.
HLMM (UHLMM) = Simple Unadjusted Change Score.
HLMM (AHLMM) = Demographic and Intake Adjusted Change Score.
MEM (LMEM) = Multivariate Layered.

clear. Value-added measures based on the SFEM are highly correlated with those based on the LMEM and could be used as a simple substitute.

Question 2:

The AHLMM (Model 3) produced value-added measures that were not consistently in close agreement with those from the SFEM (Model 1). Correlations ranged from 0.60 to 0.96 with more than half being less than 0.80 for the six grade-by-subject combinations. This result indicates that including student or school level variables, or employing shrinkage, in the AHLMM value-added assessment produces value-added measures that are notably different from those produced by the SFEM. Thus, we cannot unequivocally recommend the use of SFEM over AHLMM. The choice between these two models depends on other considerations that will be discussed in Section 8.

Question 3:

There was not consistently strong agreement of results between the multivariate LMEM (Model 4) and the AHLMM (Model 3). The correlations ranged from 0.57 to 0.94 with all but one of the six being less than 0.75. Again, it is possible that the discrepancy in results from the LMEM and AHLMM will be less when analyzing three or more times. This, however, should not be assumed without proof from future studies. In the case where only two years of data are to be used, our results suggest clearly that the choice of models, among the four considered here, can be restricted to SFEM or AHLMM. The results relevant to the secondary questions follow:

Question 4:

Shrinkage by itself had little impact on the value-added assessment of school performance. The results from Model 1 were in strong agreement with those from Model 2, with correlations ranging from 0.97 to 1.00.

Question 5:

Multivariate analysis also had little impact on the assessment of school performance. The results of Model 2 were in strong agreement with those of Model 4, with correlations ranging from 0.94 to 0.99.

Question 6:

The effect of inclusion of student and school level covariates in Model 3 had a notable impact on value-added assessment of school performance. This was reflected by the relatively weak agreement of results from Model 3 and Model 2, which differed only by the inclusion of several such covariates in Model 3. Correlations ranged from 0.61 to 0.95, with all but two being less than 0.80.

7. Example Grading System

Once a methodology has been chosen for calculating measures of value-added to students by each school to be graded ("knowledge added" may be a better term),

those measures can be translated into grades, preferably for each subject and grade level combination. The grades then could be aggregated over subjects to produce a performance summary for the team of teachers at each grade within each school, averaged over grades to produce a performance summary for math and reading teachers separately, and averaged over subjects and grades to measure the overall performance of the school. In this section, we illustrate how the aggregation could be accomplished in a naturally appealing way using grade point averaging.

Standardized value-added measures (i.e., z scores) were calculated by dividing each BLUP of random school effects by its standard error. Grades were then assigned as follows:

1. If $z > 2$, then assign a grade of A and 4 growth points;
2. If $1 < z \leq 2$, then assign a grade of B and 3 growth points;
3. If $-1 < z \leq 1$, then assign a grade of C and 2 growth points;
4. If $-2 < z \leq -1$, then assign a grade of D and 1 growth point;
5. If $z \leq -2$, then assign a grade of F and 0 growth points.

Such grading resulted in the GPAs given in Table 6. The grades assigned to each school based on Models 1 and 4 were almost identical. The grades assigned under Model 3, however, were notably different from those assigned based on Models 1 or 4. This difference can be attributed to the fact that Model 3 adjusts for sociodemographic variables when assessing school effectiveness. When using Model 3 for grading the schools, schools that had high percentages of students in "high risk" sociodemographic groups graded higher than when either Model 1 or 4 was used. On the other hand, schools that had lower percentages of such students tended to grade lower under Model 3 than under Model 1 or 4.

Variation, within and among schools, with respect to percent minority and poverty status (summary statistics for the two most important predictors in Model 3), was shown in Table 2. An investigation of the relationship of within and among school heterogeneity of GPA in Table 6 to within and among heterogeneity of percent poverty and percent minority in Table 2, although beyond the scope of this paper, might solidify the interested reader's understanding of why Model 3 results differed from those of Models 1 and 4.

The difference in grades assigned to schools based on Models 3 and 4 was manifest by generally lower correlations between the corresponding value-added measures. Whether a grading scheme should be based on Model 1 or Model 3 depends on whether schools should be held accountable for the effects of factors related to the sociodemographic make-up of their student populations.

It should be noted that our choice of cut-offs for defining grade categories was arbitrary. Other cut-offs could be chosen and GPAs calculated as in the current example.

8. Discussions and Conclusions

It is widely accepted among educators and researchers that value-added assessment of school performance is better than an assessment based on status-scores alone.

TABLE 6

Growth Point Averages for Each School Based on Value-Added Measures from Each of Three Models

	Model 1 (SFEM)						Model 3 (Adjusted HLMM)						Model 4 (LMEM)					
School	M	R	3	4	5	T	M	R	3	4	5	T	M	R	3	4	5	T
1	1.0	2.0	1.5	2.5	0.5	1.5	2.0	2.3	3.0	2.0	1.5	2.2	1.0	1.7	1.5	2.0	0.5	1.3
2	2.3	2.0	2.0	2.0	2.5	2.2	2.3	2.0	2.0	2.0	2.5	2.2	2.3	2.0	2.0	2.0	2.5	2.2
3	1.7	1.3	1.0	1.5	2.0	1.5	2.0	1.7	1.5	2.0	2.0	1.8	2.0	1.3	1.0	2.0	2.0	1.7
4	3.0	2.0	2.5	3.0	2.0	2.5	2.7	2.0	2.0	3.0	2.0	2.3	2.3	1.7	1.5	2.5	2.0	2.0
5	1.3	2.0	1.5	2.0	1.5	1.7	1.7	2.0	1.5	2.0	2.0	1.8	1.7	2.0	1.5	2.0	2.0	1.8
6	1.7	1.3	0.0	2.5	2.0	1.5	2.3	1.7	1.5	2.5	2.0	2.0	1.7	1.3	0.0	2.5	2.0	1.5
7	3.0	2.0	4.0	1.0	2.5	2.5	2.3	1.7	2.0	1.5	2.5	2.0	3.0	2.0	3.5	1.5	2.5	2.5
8	0.3	1.3	1.0	0.5	1.0	0.8	1.0	2.0	2.0	0.5	2.0	1.5	0.3	2.0	1.0	1.0	1.5	1.2
9	2.3	1.3	1.0	3.0	1.5	1.8	2.7	2.3	2.5	2.5	2.5	2.5	2.0	1.3	1.5	2.5	1.0	1.7
10	0.7	1.7	0.5	1.0	2.0	1.2	2.0	2.0	2.0	2.0	2.0	2.0	0.7	2.0	0.5	1.5	2.0	1.3
11	2.0	1.0	1.5	2.5	0.5	1.5	2.3	2.0	2.5	2.0	2.0	2.2	2.0	1.7	1.5	2.5	1.5	1.8
12	1.3	1.3	0.5	2.0	1.5	1.3	1.0	1.7	1.5	1.5	1.0	1.3	1.3	1.7	0.5	2.0	2.0	1.5
13	2.7	2.0	2.0	2.0	3.0	2.3	2.0	1.7	2.0	2.0	1.5	1.8	2.7	2.0	2.0	2.0	3.0	2.3
14	3.3	2.7	4.0	2.0	3.0	3.0	2.3	2.3	3.0	2.0	2.0	2.3	3.0	2.7	3.5	2.0	3.0	2.8
15	3.5	3.0	4.0	2.5	.	3.3	2.5	3.0	3.5	2.0	.	2.8	3.0	3.0	4.0	2.0	.	3.0
16	2.3	3.0	3.0	1.5	3.5	2.7	1.7	1.7	1.5	1.5	2.0	1.7	2.0	2.3	2.5	1.5	2.5	2.2
17	2.7	2.7	3.0	3.0	2.0	2.7	2.3	2.3	2.5	2.5	2.0	2.3	2.7	2.3	2.5	3.0	2.0	2.5
18	3.0	2.3	2.0	2.5	3.5	2.7	2.0	2.0	1.5	2.0	2.5	2.0	3.3	2.3	2.0	2.5	4.0	2.8
19	3.0	1.5	3.0	1.5	.	2.3	2.0	1.0	1.5	1.5	.	1.5	3.0	2.0	3.0	2.0	.	2.5
20	1.0	3.0	2.5	2.0	1.5	2.0	1.7	2.3	1.5	2.5	2.0	2.0	1.7	2.7	2.0	2.5	2.0	2.2
21	2.3	2.7	3.0	1.5	3.0	2.5	1.7	2.0	2.0	1.5	2.0	1.8	2.3	2.0	2.5	1.0	3.0	2.2
22	3.0	2.5	3.0	2.5	.	2.8	2.5	2.0	2.0	2.5	.	2.3	3.0	2.5	3.0	2.5	.	2.8

Note.
M = Math GPA, averaged over grades.
R = Reading GPA, averaged over grades.
T = Total GPA, averaged over grades and subjects.
3 = Third grade GPA, averaged over subjects.
4 = Fourth grade GPA, averaged over subjects.
5 = Fifth grade GPA, averaged over subjects.

Several conclusions can be drawn from the results of the current study concerning four models for value-added assessment of elementary school performance.

First, the simplest model, that is the SFEM (Model 1), is preferred over the much more complex LMEM (Model 4). These models produced results that were highly correlated ($r > 0.91$) for all elementary school grades by subject areas. Thus, there is little or no benefit to using the more complex model.

Second, the SFEM (Model 1) cannot be recommended unequivocally over the AHLMM (Model 3). The choice between these two models must be based primarily on non-empirical considerations. The crux of the issue is whether schools should be held accountable for significant effects of sociodemographic factors. If these variables are included in the model, then, in effect, schools are excused from responsibility for their effects. It is likely that schools are partially, but not wholly, responsible for such effects and should not be totally excused. Neither should they be held totally responsible. Unfortunately, the choice between Models 1 and 3 leads to taking one or the other of these undesirable positions. Model 1 might be preferred in a low-stakes accountability system that provides incentives and resources for "less effective" schools to improve and that does not base salary raises on the value-added measures. In a high stakes system, however, where teachers' salaries and school budgets depend on "high performance," not adjusting for significant sociodemographic factors could encourage the flight of good teachers and administrators from schools with high percentages of poor or minority students. On the other hand, adjusting for these factors could institutionalize low expectations for poor or minority students and thereby limit their opportunity to achieve their full potential.

Third, the isolated effects of shrinkage and multivariate analysis were not notable, while the effect of including significant student and school level covariates was. Concerning the latter point it should be noted that, if schools are partly but not wholly responsible for the effects of covariates, then bias results from either including or excluding them. Assuming partial responsibility, the exclusion of student and school level covariates from our analyses produced a bias against schools with an over representation of, for example, poverty and minority students. On the other hand, if schools were at least partially responsible for the effects of these covariates, then including them resulted in value-added measures that were biased against schools with under representation of minority or poverty students.

It was not surprising that the value-added measures produced by the AHLMM (Model 3) and the LMEM (Model 4) were not in consistent agreement, as the complexities of each model stem from attempts to improve on the simplest value-added measure (Model 1) in different ways. A common appeal of both types of mixed models is that they produce shrunken estimates of school effects. Beyond that, however, they target different perceived deficiencies of the simplest method. Similarly, the lack of agreement between the value-added measures obtained from the UHLMM (Model 2) and the AHLMM (Model 3) was not surprising. The strong agreement of results from Models 1, 2, and 4, on the other hand, was unexpected. These results suggest that the theoretical deficiencies of Model 1 that are targeted

by Models 2 or 4 have little practical impact when two years of data are used to grade schools.

Regardless of the methods chosen for value-added assessment, it is preferable to hold each school accountable for each subject by grade combination separately. By using methods like the Growth Point Average grading system presented in this paper, the results can be aggregated across all grades and subjects to obtain an overall GPA to serve as a value-added measure for the entire school.

It should be noted that the GPA grading system presented here for illustrative purposes mixes the effect sizes and precision of the value-added estimates. This means that a large school could receive a higher grade than a small school in spite of having a lower value-added measure, or a lower grade than a small school in spite of a higher value-added measure. Whether this is unfair, however, is debatable. It could be argued that grades should be based on statistical inferences about true effects and that those inferences appropriately incorporate sample size differences. Nevertheless, if this feature is unacceptable in practice, then the denominator of the z-scores could be replaced by the appropriate variance component estimate (i.e., that for u_{is2}), which is the same for all schools.

Additional work is needed to answer the questions considered in this study when developing teacher accountability systems or school accountability systems in other districts or at higher grade levels. It is not clear whether the conclusions drawn from this study of schools in a single Florida school district will generalize to teachers or to other districts. Shrinkage estimators, for example, may be quite valuable in the analysis of teacher's performance. Sanders, et al. stated, "Shrinkage estimators of teacher effects provide protection against fortuitous misclassifications of individuals" (Sanders, Saxton, & Horn, 1997). It also is not guaranteed that our conclusions will apply to more general settings where three or more years of data are available for analysis. McCaffrey, et al. (2003), when studying the estimation of teacher effects from longitudinal data, argued that allowing for cross-time correlations in the model might mitigate the effects of omitting some covariates. If so, then our use of just two years of data would undervalue a theoretical advantage of the LMEM compared with the SFEM. Additional research is needed to determine whether the LMEM would produce notably different results from the SFEM when more than two years of data are analyzed.

Notes

[1] Throughout this discussion, we loosely refer to models as producing value-added measures. It is left implicit that references to models herein are referring the model coupled with the estimation procedure used by SAS to estimate the model's parameters. The model and estimation procedure together produce value-added measures.

[2] We say "perceived theoretical deficiency" because the shrinkage estimators available for use in practice are only approximations of theoretically superior Stein-type shrinkage estimators (Efron & Morris, 1975; Morris, 1983) and the related best linear unbiased predictors (BLUPs), which are Empirical Bayes estimators (Vonesh & Chinchilli, 1996).

Appendix
Model Building Strategy for Model 3

The model building strategy presented in this Appendix was applied to all available data to build an HLMM that included subsets of potentially important student and school level variables for each subject area by grade separately. The student-level variables initially considered for inclusion were: intake score, minority status (yes, no), gender (male, female), poverty status (poverty, non-poverty), retention status (retained, not retained), and mobility (transfer from other school: yes, no). The school-level variables considered initially were: mean intake score for the particular grade by school by subject area, percent minorities in the school, percent males in the school, percent poverty students in the school, percent of students in the school who had changed schools since the test in 1998, total school enrollment, and average class size, and the school-wide percentage of students retained. The models were fitted to student specific differences (d) between ITBS achievement scale scores in 1999 ($t = 2$) and 1998 ($t = 1$).

A two-step procedure was used to specify and fit the AHLMM (Model 3). A stage-wise, step-wise backward selection process was used in the model building strategy to obtain the final model specification. Then, the specified model was fitted. Details of these two steps follow:

1. School effects were initially treated as fixed, and ordinary least squares (OLS) were used to determine which student and school-level variables to include in the model. The final OLS model was obtained using a two-stage backward selection procedure. At stage one, the starting model included all potentially important student-level variables, their two- and three-way interactions, and a school factor and its interactions with the student level variables and interactions. Blocks of interactions associated with each student level variable (e.g., all interactions involving the poverty variable) were tested one at a time by deleting them from the full model. The block with the largest p-value for the full and reduced model F-statistic was dropped at each step until all remaining blocks in the model were significant. Then, term-by-term backward selection was used to delete nonsignificant terms. Interactions were dropped from the model before any included main effects. At the end of Stage 1, only student-level covariates with corresponding p-values smaller than 0.01 were left in the model.

   At Stage 2, the school factor was replaced by a list of all potentially important school level covariates, producing a model with the school-level variables and their interactions with the student-level covariates that interacted with the school factor in Stage 1. Backward selection was again used to drop nonsignificant school level covariates or interactions from the model using a cutoff alpha level of 0.01. SAS PROC GLM was used to implement the model fitting in this step.

2. The final model obtained was then refit as an HLMM assuming random school effects in order to obtain shrinkage estimates. Random school effects were included for each significant school effect (main or interaction) left in the model after stage 1 of the model building procedure. For example, if a school-by-poverty

interaction was significant in the final model in Stage 1, then the poverty variable was assumed to have a random coefficient in the AHLMM. SAS PROC MIXED was used to fit the model suggested by the results from Step 1. The default estimation method for the variance components in SAS PROC MIXED was restricted maximum likelihood (REML), and the SAS procedure automatically provided shrinkage estimates of the school effects (*i.e.,* BLUPs). This default was used throughout.

Table 4 contains the estimated coefficients on variables that had significant effects on learning gains for each grade by subject combination. In all cases, for the elementary schools in this particular county, the only random coefficients that were identified by the model building procedure were random intercepts.

These results are presented to show which student or school level variables had significant effects on learning gains and to show the magnitude of their effects. This may be important information for policy makers when determining which, if any, such variables to include in a value-added accountability system.

References

Aitkin, M., & Longford, N. (1986), Statistical modeling in school effectiveness studies. *Journal of the Royal Statistical Society, A, 149,* 1–43.

Carter, R. L., et al. (2001). *Annual learning gains/value-added methods for assessing student achievement, teacher effectiveness, and school accountability: Value-added analysis models for assessing school effectiveness using simple gain (Model II.A), fixed effects gain (Model III.A) and HLM (IV.A) approaches based on data from three school districts* (Project # 011-90950-00004), Report to the Florida Department of Education Assessment, Testing and Evaluation Services.

Coleman, J. S., Campbell, T. E., & Kilgore, S. B. (1982). *High achievement: Public, Catholic, and other private schools compared,* New York: Basic.

Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its Generalizations. *Journal of the American Statistical Association, 74,* 311–19.

Goldstein, H. (1997). Methods in school effectiveness research. *School effectiveness and school improvement, 8*(4), 369–395.

Laird, N., & Ware, J. R. (1982). Random effect models for longitudinal data. *Biometrics, 38,* 963–974.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *Preface to SAS system for mixed models,* Cary, NC: SAS Institute Inc.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., Hamilton, L., & Kirby, S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–102.

Morris, C. (1983). Parametric empirical Bayes inference, theory and applications. *Journal of the American Statistical Association, 78,* 47–65.

Olson, L. (1998). A question of value. *Education Week On the Web, 17*(35), Retrieved February 2, 2004 from http://www.edweek.org/ew//vol-17/35value.h17.

Phillips, G. W., & Adcock, E. P. (1996, April). *Practical applications of hierarchical linear models to district evaluations.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59,* 1–17.

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,* 299–311.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Educational Assessment System (TVAAS): 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, Inc.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*(4), 323–355.

Vonesh, E. F., & Chinchilli, V. M. (1997). *Linear and nonlinear models for the analysis of repeated measurements,* New York: Marcel Dekker, Inc.

Authors

CARMEN D. TEKWE is Statistician, Center on Aging and Health, Johns Hopkins University, 2024 E. Monument Street Suite 2-700 Baltimore, Maryland, 21212; ctekwe1@jhmi.edu. Her areas of specialization are longitudinal studies and linear and nonlinear mixed-effect models.

RANDY L. CARTER is Professor, Department of Biostatistics, University at Buffalo, Farber Hall Rm. 249, 3435 Main St., Buffalo, NY 14214-3000 rcarter@buffalo.edu. His area of specialization is statistics.

CHANG-XING MA is Research Assistant Professor, Department of Statistics, University of Florida, Box 100212 Gainesville, FL, 32610-0212; cma@biostat.ufl.edu. His areas of specialization are experimental designs, statistical genetics and public health.

JAMES ALGINA is Professor, Department of Measurement and Evaluation, University of Florida, PO Box 117047, Gainesville, FL 32611-7047; algina@ufl.edu. His areas of specialization are applied statistics and psychometric therapy.

MAURICE E. LUCAS is Director, Office of Research and Evaluation, Alachua County School Board, 620 East University Avenue Gainesville, FL 32601; lucasme@sbac.ufl.edu. His areas of specialization are administration of education testing programs, program evaluation, and research.

JEFFREY ROTH is Associate Professor, Department of Pediatrics, University of Florida and Associate Director of the Maternal Child Health and Education Research and Data Center, Box 100296 Gainesville, FL 32610-0296; jeffroth@ufl.edu. His area of specialization is program evaluation.

MARIO ARIET is Professor, Department of Medicine, University of Florida and Associate Director of the Maternal Child Health and Education Research and Data Center, Box 100372, Gainesville, FL 32610-0372; arietm@medcs.ufl.edu. His area of specialization is computer science.

THOMAS FISHER is President, Fisher Education Consulting, Inc., 555 Hickory Blvd, McMinnville, TN 37110; thfisher@blomand.net. His areas of specialization are large-scale educational assessment and school accountability.

MICHAEL B. RESNICK is Professor, Department of Pediatrics, University of Florida, and Director of the Maternal Child Health and Education Research and Data Center, Box 100296 Gainesville, FL 32610-0296; mresnick@ufl.edu. His areas of specialization are educational psychology, child development, maternal child health, and educational outcomes research.

Manuscript Received February 2003
Revision Received October 2003
Accepted December 2003