Calculating Missing Student Data in Hierarchical Linear Modeling: Uses and Their Effects on School Rankings

Timothy H. Orsak Robert L. Mendro Dash Weerasinghe
Dallas Public Schools

In the age of student accountability, public school systems must find procedures for identifying effective schools, classrooms and teachers that help students continue to excel academically. As a result, researchers have been modeling schools to calculate achievement indicators that will withstand not only statistical review but political criticism. One of the numerous issues encountered in modeling is the management of missing student data. This paper addresses three techniques that elucidate the effects of absent data and highlight consequences on school achievement indicators. The outcomes of each technique are estimated data and School Effectiveness Indices (SEIs). A set of criteria is established from an original data set to determine a baseline to which the analyses will be compared in determining the most appropriate approach in estimating missing data..

Completeness of any data base should be considered a rarity when managing educational data. Numerous factors, not limited to student lack of attendance, data misinterpretation, and mistakes in data entry, all affect the accuracy of any educational database. While incorrect data scores are difficult, if not impossible, to detect, missing scores are readily identifiable. Effective schools within the Dallas Public Schools have been identified by statistical methodologies for several years. Many years of analyses have deduced the accuracy of statistical methods' rankings of schools within the district. Yet these analyses utilized only student data that was complete for both post-test and pre-test years. On average, between 8% and 12% of student data cannot be included in yearly calculations due to at least one year of missing test scores. However, attempts to use all available data while not introducing extraneous trends could more accurately help identify effective schools. In this paper, the question of best estimation of absent post-test data is addressed.

The current problem faced in the computation of school effectiveness rankings relates to missing student test data. How could we effectively rank the school of interest without complete data for its constituents? Several publications have addressed treatment of missing scores in data sets through the use of inference, replacement of missing values with probable values, etc. One example is Sanders, et.al. (1993), which implemented a sparse matrix mixed modeling program to predict missing student values. Yet with the typical school district not having the resources to implement such a program, what would be the most effective and efficient method for school analysis? Dallas Public Schools has addressed the missing data issue by not including it in any analysis, thus eliminating possible influences.

The analysis comprised of 5197 6th grade students who had complete raw data scores for the *Iowa Test of Basic Skills* mathematics and reading tests for years 1995 and 1996 and student characteristics of ethnicity, English proficiency status, census poverty data, census college data, and gender. To analyze the effects of missing data, specific percentages of the post-test scores from the original data set were randomly deleted which produced reduced data sets. The percentages of data deleted in this study were 1%, 2%, 5%, 10%, and 20%. The reduced data sets were then evaluated by *Scientific Software's* HLM2L hierarchical linear modeling software and by MicroSofts' Excel's Ordinary Least Squares software program to produce regression coefficients for each school. The deleted post-test scores were then estimated by HLM, by OLS and by the average post-test score per school. The three new data sets composed of HLM estimates of missing data, OLS estimates of missing data, and average post-test data per school and the original data set (non-deleted scores), were then reprocessed by HLM and school effectiveness indices (SEIs) generated. The SEIs were calculated from HLM as the estimated Bayesian (EB) residuals for the school level intercept rescaled to a mean of 50 and standard deviation of 10. The EB residual reflects the overall achievement of the students within a school. The SEIs from the new data sets were compared to the original data set's SEI scores whereas the estimated post-test scores were compared to the actual scores that were deleted. This process was carried out for three models of varying complexity.

Investigation and Procedure

This study expands previous studies of HLM to investigate the effects of missing data through the use of HLM models in ranking 118 elementary schools from the Dallas Public Schools at the sixth grade *(Webster, et. al., 1994, 1995; Mendro, et. al., 1994, 1995; Orsak, et. al., 1996)*. Ten school characteristics variables were available for each school. To eliminate undue influences from varying school sizes, the original 5197 student data set was randomly reduced such that exactly 30 students were included per school. This created a new, reduced data file which contained 2610 students within 87 schools. Initial analyses for this reduced data set explored OLS and HLM estimates from three models, each more complex than the previous. Then all 5197 students were used in a fourth analysis. The initial exploratory analysis involved simple data analysis for the reduced data set. ****

Table 1. Student Characteristic Correlations

	GEN	*LUN*	*BLK*	*HIS*	*LEP*	*INC*	*POV*	*COL*	*R-95*	*M-95*	*M-96*
GEN	1.000										
LUN	-.0122	1.000									
BLK	.0138	.1112	1.000								
HIS	-.0278	.0827	-.6043	1.000							
LEP	.0193	.1390	-.3049	-.1806	1.000						
INC	-.0090	.3407	.2046	.0418	.0215	1.000					
POV	-.0253	.2903	.1530	.0236	.0634	.5804	1.000				
COL	-.0172	.3461	-.0143	.2433	.1412	.6135	.3453	1.000			
R-95	.0951	.2282	.1992	-.0997	.1086	.1863	.1369	.2061	1.000		
M-95	.0169	.1747	.1451	-.0750	.0907	.1682	.1220	.1761	.6112	1.000	
M-96	.0354	.1763	.1303	-.0522	.0966	.1566	.1131	.1901	.5605	.7857	1.000

** GEN is Gender, LUN is Free Lunch Status, BLK represents Black, HIS represents Hispanic, LEP is Limited English Proficient, INC is average block income, POV is percent block poverty, COL is percent block college, R-95 is ITBS Reading for 1995, M-95 is ITBS Mathematics for 1995, M-96 is ITBS Mathematics for 1996.

Table 2. Student Characteristic Summary

	N	*MEAN*	*SD*	*MIN*	*MAX*
GEN	2610	1.54	.50	1	2
LUN	2610	1.28	.45	1	2
BLK	2610	1.50	.5	1	2
HIS	2610	1.74	.44	1	2
LEP	2610	1.92	.28	1	2
INC	2610	28139.44	14488.61	1290	185017.00
POV	2610	74.73	20.88	0	100
COL	2610	9.15	13.12	0	100
R-95	2610	11.91	4.42	1	22
M-95	2610	34.95	8.66	11	54
M-96	2610	37.83	9.23	9	59

** See Table 1 Legend

The models used for the prediction of deleted post-test data are as follows. Analyses began with a basic model for prediction and increased in complexity.

The models with no student level variables and no school level variables:

Model 1A (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$

$$\beta_{1k} = \gamma_{10} + u_{1k}$$

****

1. It must be value-added.

2. It must include multiple outcome variables.

3. Schools must only be held accountable for students who have been exposed to their instructional program (continuously enrolled students).

4. It must be fair. Schools must derive no particular advantage by starting with high-scoring or low-scoring students, minority or white students, high or low socioeconomic level students, or limited English proficient or non-limited English proficient students. In addition such factors as student mobility, school overcrowding, and staffing patterns over which the schools have no control must be taken into consideration.

5. It must be based on cohorts of students, not cross-sectional data.

Within the five aforementioned parameters, a number of statistical models are possible. The two most widely cited approaches in the literature involve various uses of basic ordinary least squares regression techniques (OLS regression) (Aiken and West, 1991; Bano, 1985; Felter and Carlson, 1985; Kirst, 1986; Klitgard and Hall, 1973; McKenzie, 1983; Millman, 1981; Saka, 1984) or the use of a variety of hierarchical linear models (HLM) (Bryk, et.al., 1988; Bryk and Raudenbush, 1992; Bryk and Thum, 1996; Dempster, Rubin and Tsutakawa, 1981; Elston and Grizzle, 1962; Goldstein, 1987; Henderson, 1984; Laird and Ware, 1982; Mason, Wong, and Entwistle, 1984; Rosenberg, 1973).

This study is the sixth in a series of studies conducted in the Dallas Independent School District over a period of eight years. All models addressed in these studies have been designed to isolate the effect of a given school's or teacher's practices on important student outcomes. The school effect is conceptualized as the difference between a given student's performance in a particular school and the performance that would have been expected if that student had attended a school with similar context but with practice of average effectiveness. The teacher effect is conceptualized similarly at the teacher level. The results of previous studies have suggested:

- Utilizing basic OLS regression models with individual student growth curves and no demographic variables produced results that were uncorrelated with student level demographic variables and slightly correlated with school level demographic variables but not with pretest levels (Webster and Olson, 1988).

- Utilizing basic OLS regression models with school level variables produced results that were unreliable and that were correlated with student level demographic variables and student level pretest scores. Too much important information is lost in this process (Mendro and Webster, 1993).

- Utilizing two stage OLS regression models, the first stage removing the effects of student demographic variables from both the pretest and posttest measures, produced results that were uncorrelated with student pretest scores and student level demographic variables and only minimally correlated with school level demographic variables (Webster, Mendro, and Almaguer, 1994). These models are discussed later in this paper.

- Utilizing student based two-stage OLS regression models that accounted for first and second order interactions among basic demographic variables produced results at the school level that were very reliable, that correlated very highly with those produced by two-stage, two level-HLM ($\geq .97$), and that were uncorrelated with student and school level demographic variables and pretest scores. It was noted, however, that adding school level variables as conditioning variables in HLM drove the correlations with school level variables to absolute zero (Webster, Mendro, Bembry, and Orsak, 1995).

- Utilizing basic unadjusted gain scores to rank schools produced results that were not highly correlated with results produced by either OLS student-level regression models or two-level HLM ($<.75$). Further, gain models produced results that were correlated with some student and school level demographic variables and with pretest score. Using straight NCE scores to rank schools produced results that correlated poorly with the results obtained from the OLS and HLM models ($<.55$) and were highly correlated with both student level and school level demographic variables as well as pretest score (Webster, Mendro, Bembry, and Orsak, 1995).

- Utilizing student based two-stage OLS regression models that accounted for first and second order interactions among basic demographic variables produced results that were very close to those produced by two-stage, two-level HLM at the school level and, when adjusted for shrinkage, produced results at the teacher level that correlated very highly with the results of two-level and three-level HLM models (≥.90). Most models accounted for more than seventy percent of the variance in student achievement in reading and mathematics and produced extremely consistent results. Correlations of results with important school, teacher, and student level contextual variables and with pre-score characteristics were negligible for all models (Webster, Mendro, Orsak, and Weerasinghe, 1996).

In a recent thought-provoking critique of the Dallas models, Thum and Bryk (1997) raised some questions that were responded to in a response to Thum and Bryk that will be published in an upcoming book on teacher evaluation that is edited by Jason Millman (1997). This study further addresses the points raised by Thum and Bryk as well as consolidates previous research by using only the best models from OLS regression and HLM for comparison purposes. The major objective of this study is to determine the most reliable and efficient methodology for identifying effective schools and teachers.

Except for the original Webster and Olson (1988) study, all other Dallas studies have used only elementary grades as their samples. There are a large number of elementary schools in the Dallas Public Schools (≥125, depending on the grade studied). This study utilizes sixth and eighth grades in an effort to ensure that the number of schools does not significantly effect the results. (There are 127 schools with sixth grades and only 26 with eighth grades.) In order to keep the study simple, the only outcome variable used is 1996 *Iowa Tests of Basic Skills Reading (ITBS)* and the only cognitive measures predictor variables are *ITBS* Reading and *ITBS* Mathematics tests. The actual system for which these equations are used includes multiple outcome and predictor variables and is described in detail in a companion paper by Webster, Mendro, Bembry, and Bearden (1997).

This study investigates a number of methodological issues related to the use of various mathematical models for estimating school and teacher effect. The Thum and Bryk (1997) concerns are addressed as well as a number of other issues related to the effectiveness of various models. The major areas of investigation include:

1. Is there any significant difference between results produced by a two-stage model as opposed to including all relevant demographic and cognitive measures in a one-stage equation? The authors have always believed that there is no practical difference. Thum and Bryk suggested that the two stage process may be less reliable because residuals from a set of residuals are unreliable.

2. Is there a significant difference between results produced by assuming random slopes versus fixed slopes at the second and third levels in HLM? This question is also related to the two-stage questions since with complex data sets one generally cannot solve many one-stage HLM models assuming random slopes. If one assumes fixed slopes, the HLM algorithms generally will solve the equations.

3. Does a three-level HLM that uses student gain scores as the outcome variable with no school level conditioning variables and limited student level conditioning variables, similar to that proposed by Bryk and Thum (1996), produce results that are comparable to those produced by similar status-based models? Status-based models are models that do not utilize gain scores as the basic unit of analysis and include all other models discussed in this paper.

4. How free from bias are the estimates relative to important school, teacher, and student contextual variables?

5. How free from bias are the estimates relative to pretest scores?

6. Given the complexity of the three-level HLM model in estimating teacher effect, particularly in terms of data requirements, can the results produced by a three-level HLM model be validly approximated through the use of a two-level HLM-model with a shrinkage adjustment?

7. Can a longitudinal student growth curve approach to predicting school and teacher effect produce bias free results without specifically addressing student, teacher, and school contextual variables?

Method

Sample

The samples used in this study consisted of all students who were enrolled and tested in the Dallas Public Schools in grade 5 in 1995 and grade 6 in 1996; in grade 7 in 1995 and grade 8 in 1996; and, in the multi-year longitudinal studies, students who were enrolled and tested in the Dallas Public Schools in grade 2 in 1992, grade 3 in 1993, grade 4 in 1994, grade 5 in 1995, and grade 6 in 1996; and in grade 4 in 1992, grade 5 in 1993, grade 6 in 1994, grade 7 in 1995, and grade 8 in 1996. All samples represent longitudinal cohorts of real students.

Instrumentation

The instrumentation used for the study was the *Iowa Test of Basic Skills* Reading and Mathematics subtests. Raw scores were the unit of analysis. Reading was the only criterion variable used.

School Effect

Fifteen different OLS regression and HLM models were investigated to determine their reliability and appropriateness for measuring school effect. Figure 1 contains descriptions of these models. The numbers used to describe the models in this section are from the numbers associated with each model in Figure 1. Model 1, for example, is Basic OLS regression as described in Figure 1. Each model was investigated in terms of its efficiency of prediction; the reliability of school ranks produced; the amount of variance accounted for; the amount of bias relative to important school, classroom and students contextual variables; and, the amount of bias relative to pretest scores. All comparisons are in terms of the effectiveness indices produced by each of the models. Correlations that appear in later comparisons in the results section are correlations between the various estimates of effect produced by the various models.

Student level variables included in a number of the OLS regression and HLM models were:

Y_{ij} = Outcome variable of interest for each student *i* in school *j*.
X_{1ij}= Black English Proficient Status (1 if black, 0 otherwise).
X_{2ij}= Hispanic English Proficient Status (1 if Hispanic, 0 otherwise).
X_{3ij}= Limited English Proficient Status (1 if LEP, 0 otherwise).
X_{4ij}= Gender (1 if male, 0 if female).
X_{5ij}= Free or Reduced Lunch Status (1 if subsidized, 0 otherwise).
X_{6ij}= Block Average Family Income.
X_{7ij}= Block Average Family Education.
X_{8ij}= Block Average Family Poverty Level.
X_{kij}= Indicates the variable *k* of *i-th* student in school *j* for $i = 1,2, ..., I_j$ and $j = 1, 2, ..., J$.

Classroom level variables included in a number of the HLM models were:

T_{1j}= Classroom Mobility.
T_{2j}= Classroom Overcrowdedness.
T_{3j}= Classroom Average Family Education.
T_{4j}= Classroom Average Family Education.
T_{5j}= Classroom Average Family Poverty Index.
T_{6j} = Classroom Percentage on Free or Reduced Lunch.
T_{7j}= Classroom Percentage Minority.
T_{8j}= Classroom Percentage Black.
T_{9j}= Classroom Percentage Hispanic.
T_{10j}= Classroom Percentage Limited English Proficient.

School level variables included in a number of the HLM models were:

W_{1j}= School Mobility.
W_{2j}= School Overcrowdedness.
W_{3j}= School Average Family Education.
W_{4j}= School Average Family Education.
W_{5j}= School Average Family Poverty Index.
W_{6j}= School Percentage on Free or Reduced Lunch.
W_{7j}= School Percentage Minority.
W_{8j}= School Percentage Black.
W_{9j}= School Percentage Hispanic.
W_{10j}= School Percentage Limited English Proficient.

Predictor and Criterion variables included in various models were:

Criterion Variables

$ITBS\_RES\_R\_96_{ij}$ =1996 ITBS Residual Reading score from fairness stage calculated as an OLS residual for student *i* in school *j*.

$ITBS\_R\_96_{ij}$ =1996 ITBS Reading Score.

$ITBS\_GAIN\_R_{96}\_R_{95\,ij}$ = ITBS Gain Score for 1995 to 1996.

Predictor Variable

$ITBS\_RES\_R\_95_{ij}$ = 1995 ITBS Residual Reading score from fairness stage calculated as an OLS residual for student *i* in school *j*.

$ITBS\_RES\_M\_95_{ij}$ = 1995 ITBS Residual Mathematics score from fairness stage calculated as an OLS residual for student *i* in school *j*.

$ITBS\_RES\_R\_94_{ij}$ = 1994 ITBS Residual Reading score from fairness stage calculated as an OLS residual for student *i* in school *j*.

$ITBS\_RES\_M\_94_{ij}$ = 1994 ITBS Residual Mathematics score from fairness stage calculated as an OLS residual for student *i* in school *j*.

$ITBS\_RES\_R\_93_{ij}$ = 1993 ITBS Residual Reading score from fairness stage calculated as an OLS residual for student *i* in school *j*.

$ITBS\_RES\_M\_93_{ij}$ = 1993 ITBS Residual Mathematics score from fairness stage calculated as an OLS residual for student *i* in school *j*.

$ITBS\_RES\_R\_92_{ij}$ = 1992 ITBS Residual Reading score from fairness stage calculated as an OLS residual for student *i* in school *j*.

$ITBS\_RES\_M\_92_{ij}$ = 1992 ITBS Residual Mathematics score from fairness stage calculated as an OLS residual for student *i* in school *j*.

$ITBS\_R\_95_{ij}$ = 1995 ITBS Reading Score for student *i* in school *j*.

$ITBS\_M\_95_{ij}$ = 1995 ITBS Mathematics Score for student *i* in school *j*.

$ITBS\_R\_94_{ij}$ = 1994 ITBS Reading Score for student *i* in school *j*.$ITBS\_M\_94_{ij}$ = 1994 ITBS Mathematics Score for student *i* in school *j*.

$ITBS\_R\_93_{ij}$ = 1993 ITBS Reading Score for student *i* in school *j*.

$ITBS\_M\_93_{ij}$ = 1993 ITBS Mathematics Score for student *i* in school *j*.

$ITBS\_R\_92_{ij}$ = 1992 ITBS Reading Score for student *i* in school *j*.

$ITBS\_M\_92_{ij}$ = 1992 ITBS Mathematics Score for student *i* in school *j*.

The comparisons of results produced by Models 1 and 2 address whether or not there are differences between the effectiveness statistics produced by basic OLS Regression and basic two-level HLM. The HLM Model assumes fixed slopes at the conditioning level since the HLM algorithms could not solve these equations if random slopes were assumed. Appropriate equations for Model 1 and 2 follow:

Model 1

$$ITBS\_R\_96_{ij} = \Lambda_0 + \Lambda_1 X_{1ij} + \Lambda_2 X_{2ij} + \Lambda_3 X_{3ij} + \Lambda_4 X_{4ij} + \Lambda_5 X_{5ij} + \Lambda_6 X_{6ij} + \Lambda_7 X_{7ij} + \Lambda_8 X_{8ij} + \Lambda_9 (X_{1ij} X_{4ij}) + \Lambda_{10} (X_{2ij} X_{4ij}) + \Lambda_{11} (X_{3ij} X_{4ij}) + \Lambda_{12} (X_{1ij} X_{5ij}) + \Lambda_{13} (X_{2ij} X_{5ij}) + \Lambda_{14} (X_{3ij} X_{5ij}) + \Lambda_{15} (X_{4ij} X_{5ij}) + \Lambda_{16} (X_{1ij} X_{4ij} X_{5ij}) + \Lambda_{17} (X_{2ij} X_{4ij} X_{5ij}) + \Lambda_{18} (X_{3ij} X_{4ij} X_{5ij}) + \Lambda_{19} ITBS\_R\_95_{ij} + \Lambda_{20} ITBS\_M\_95_{ij} + \varepsilon_{ij}$$

$$SEI_j = \frac{\sum_{i=1}^{N_j} \varepsilon_{ij}}{N_j}$$

Model 2

Level 1:

$$ITBS\_R\_96_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \beta_{3j} X_{3ij} + \beta_{4j} X_{4ij} + \beta_{5j} X_{5ij} + \beta_{6j} X_{6ij} + \beta_{7j} X_{7ij} + \beta_{8j} X_{8ij} + \beta_{9j} (X_{1ij} X_{4ij}) + \beta_{10j} (X_{2ij} X_{4ij}) + \beta_{11j} (X_{3ij} X_{4ij}) + \beta_{12j} (X_{1ij} X_{5ij}) + \beta_{13j} (X_{2ij} X_{5ij}) + \beta_{14j} (X_{3ij} X_{5ij}) + \beta_{15j} (X_{4ij} X_{5ij}) + \beta_{16j} (X_{1ij} X_{4ij} X_{5ij}) + \beta_{17j} (X_{2ij} X_{4ij} X_{5ij}) + \beta_{18j} (X_{3ij} X_{4ij} X_{5ij}) + \beta_{19j} ITBS\_R\_95_{ij} + \beta_{20j} ITBS\_M\_95_{ij} + \delta_{ij}$$

where

$$\delta_{ij} \overset{iid}{\sim} N(0,\sigma^2).$$

Level 2:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{kj} = \gamma_{k0} \text{ for } k = 1, 2, ..., 20$$

$$E[u_{0j}] = 0,\ Var[u_{0j}] = \tau^2, \text{ and } u_{0j} \perp \delta_{ij}$$

$$SEI_j^* = u_{0j}^*$$

Models 3, 4, and 5 address a number of issues. First, the comparison of the results obtained from Models 1 and 3, as well as Models 2 and 4, will begin to address the one-stage versus two-stage issue. (This issue will be further addressed when the indices produced by Models 7 and 8 as well as Models 11 and 12 are compared.) The comparison of the results produced by Models 4 and 5 will address the fixed versus random slopes issue. Appropriate equations for Models 3, 4, and 5 are as follows:

Model 3

STAGE 1:

$$Y_{ij} = \Lambda_0 + \Lambda_1 X_{1ij} + \Lambda_2 X_{2ij} + \Lambda_3 X_{3ij} + \Lambda_4 X_{4ij} + \Lambda_5 X_{5ij} + \Lambda_6 X_{6ij} + \Lambda_7 X_{7ij} + \Lambda_8 X_{8ij} + \Lambda_9 (X_{1ij} X_{4ij}) + \Lambda_{10}(X_{2ij} X_{4ij}) + \Lambda_{11}(X_{3ij} X_{4ij}) + \Lambda_{12}(X_{1ij} X_{5ij}) + \Lambda_{13}(X_{2ij} X_{5ij}) + \Lambda_{14}(X_{3ij} X_{5ij}) + \Lambda_{15}(X_{4ij} X_{5ij}) + \Lambda_{16}(X_{1ij} X_{4ij} X_{5ij}) + \Lambda_{17}(X_{2ij} X_{4ij} X_{5ij}) + \Lambda_{18}(X_{3ij} X_{4ij} X_{5ij}) + r_{ij}$$

where Y_{ij} is ITBS_R_96$_{ij}$, ITBS_R_95$_{ij}$, and ITBS_M_95$_{ij}$. These will produce ITBS_RES_R_96$_{ij}$, ITBS_RES_R_95$_{ij}$, and ITBS_RES_M_95$_{ij}$, respectively.

STAGE 2:

$$\text{ITBS\_RES\_R\_96}_{ij} = \beta_0 + \beta_1 \text{ITBS\_RES\_R\_95}_{ij} + \beta_2 \text{ITBS\_RES\_M\_95}_{ij} + \varepsilon_{ij}$$

$$\text{SEI}_j = \frac{\sum_{i=1}^{N_j} \varepsilon_{ij}}{N_j}$$

Model 4

STAGE 1:

$$Y_{ij} = \Lambda_0 + \Lambda_1 X_{1ij} + \Lambda_2 X_{2ij} + \Lambda_3 X_{3ij} + \Lambda_4 X_{4ij} + \Lambda_5 X_{5ij} + \Lambda_6 X_{6ij} + \Lambda_7 X_{7ij} + \Lambda_8 X_{8ij} + \Lambda_9 (X_{1ij} X_{4ij}) + \Lambda_{10}(X_{2ij} X_{4ij}) + \Lambda_{11}(X_{3ij} X_{4ij}) + \Lambda_{12}(X_{1ij} X_{5ij}) + \Lambda_{13}(X_{2ij} X_{5ij}) + \Lambda_{14}(X_{3ij} X_{5ij}) + \Lambda_{15}(X_{4ij} X_{5ij}) + \Lambda_{16}(X_{1ij} X_{4ij} X_{5ij}) + \Lambda_{17}(X_{2ij} X_{4ij} X_{5ij}) + \Lambda_{18}(X_{3ij} X_{4ij} X_{5ij}) + r_{ij}$$

STAGE 2:

Level 1:

$$\text{ITBS\_RES\_R\_96}_{ij} = \beta_{0j} + \beta_{1j} \text{ITBS\_RES\_R\_95}_{ij} + \beta_{2j} \text{ITBS\_RES\_M\_95}_{ij} + \delta_{ij}$$

where

$$\delta_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

Level 2:

$\beta_{kj} = \gamma_{k0} + u_{kj}$ for $k = 0, 1, 2$,

where $E[u_{kj}] = 0$, $\text{Var-Cov}[u_{kj}] = T$, and $u_{kj} \perp \delta_{ij}$.

$$\text{SEI}_j^* = u_{0j}^*$$

Model 5

STAGE 1:

$$Y_{ij} = \Lambda_0 + \Lambda_1 X_{1ij} + \Lambda_2 X_{2ij} + \Lambda_3 X_{3ij} + \Lambda_4 X_{4ij} + \Lambda_5 X_{5ij} + \Lambda_6 X_{6ij} + \Lambda_7 X_{7ij} + \Lambda_8 X_{8ij} + \Lambda_9 (X_{1ij} X_{4ij}) + \Lambda_{10}(X_{2ij} X_{4ij}) + \Lambda_{11}(X_{3ij} X_{4ij}) + \Lambda_{12}(X_{1ij} X_{5ij}) + \Lambda_{13}(X_{2ij} X_{5ij}) + \Lambda_{14}(X_{3ij} X_{5ij}) + \Lambda_{15}(X_{4ij} X_{5ij}) + \Lambda_{16}(X_{1ij} X_{4ij} X_{5ij}) + \Lambda_{17}(X_{2ij} X_{4ij} X_{5ij}) + \Lambda_{18}(X_{3ij} X_{4ij} X_{5ij}) + r_{ij}$$

STAGE 2:

Level 1:

$$\text{ITBS\_RES\_R\_96}_{ij} = \beta_{0j} + \beta_{1j} \text{ITBS\_RES\_R\_95}_{ij} + \beta_{2j} \text{ITBS\_RES\_M\_95}_{ij} + \delta_{ij}$$

Level 2:

$\beta_{0j} = \gamma_{00} + u_{0j}$

$\beta_{kj} = \gamma_{k0}$ for $k = 1, 2$.

$E[u_{0j}] = 0$, $\text{Var}[u_{0j}] = \tau^2$, and $u_{0j} \perp \delta_{ij}$

$$\text{SEI}_j^* = u_{0j}^*$$

Models 6, 7, and 8 move the comparisons to a higher level of sophistication. Utilizing full models proven in previous studies, the Model 6 versus Model 7 comparison again addresses the fixed versus random slopes issue. The choice of fixed versus random slopes depends on the investigators' beliefs about the sources of variation in the slopes. The slopes are modeled using a number of school parameters at the second level. In the full model these include the school level variables listed under the conditioning variables column in Figure 1. To the extent that slopes vary as a result of these factors, their use adjusts the differences. Under these circumstances, a random model would control for the effects of possible interactions of concomitant variables in specific school settings. If there was evidence of an interaction of school effect with the conditioning variables, the fixed model would be preferable since the use of a random model would mask these effects. The Model 8 comparison with the results of Model 7 addresses the one versus two-stage issue. Appropriate equations for Models 6, 7 and 8 are as follows:

Model 6

STAGE 1:

$$Y_{ij} = \Lambda_0 + \Lambda_1 X_{1ij} + \Lambda_2 X_{2ij} + \Lambda_3 X_{3ij} + \Lambda_4 X_{4ij} + \Lambda_5 X_{5ij} + \Lambda_6 X_{6ij} + \Lambda_7 X_{7ij} + \Lambda_8 X_{8ij} + \Lambda_9(X_{1ij}X_{4ij}) + \Lambda_{10}(X_{2ij}X_{4ij}) + \Lambda_{11}(X_{3ij}X_{4ij}) + \Lambda_{12}(X_{1ij}X_{5ij}) + \Lambda_{13}(X_{2ij}X_{5ij}) + \Lambda_{14}(X_{3ij}X_{5ij}) + \Lambda_{15}(X_{4ij}X_{5ij}) + \Lambda_{16}(X_{1ij}X_{4ij}X_{5ij}) + \Lambda_{17}(X_{2ij}X_{4ij}X_{5ij}) + \Lambda_{18}(X_{3ij}X_{4ij}X_{5ij}) + r_{ij}$$

STAGE 2:

Level 1:

$$ITBS\_RES\_R\_96_{ij} = \beta_{0j} + \beta_{1j} ITBS\_RES\_R\_95_{ij} + \beta_{2j} ITBS\_RES\_M\_95_{ij} + \delta_{ij}$$

Level 2:

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1} W_{1j} + \gamma_{k2} W_{2j} + \ldots + \gamma_{k10} W_{10j} + u_{kj}$$

for $k = 0, 1, 2$.

$E[u_{kj}] = 0$, $\text{Var-Cov}[u_{kj}] = T$, and $u_{kj} \perp \delta_{ij}$

$$SEI_j^* = u_{0j}^*$$

Model 7

STAGE 1:

$$Y_{ij} = \Lambda_0 + \Lambda_1 X_{1ij} + \Lambda_2 X_{2ij} + \Lambda_3 X_{3ij} + \Lambda_4 X_{4ij} + \Lambda_5 X_{5ij} + \Lambda_6 X_{6ij} + \Lambda_7 X_{7ij} + \Lambda_8 X_{8ij} + \Lambda_9(X_{1ij}X_{4ij}) + \Lambda_{10}(X_{2ij}X_{4ij}) + \Lambda_{11}(X_{3ij}X_{4ij}) + \Lambda_{12}(X_{1ij}X_{5ij}) + \Lambda_{13}(X_{2ij}X_{5ij}) + \Lambda_{14}(X_{3ij}X_{5ij}) + \Lambda_{15}(X_{4ij}X_{5ij}) + \Lambda_{16}(X_{1ij}X_{4ij}X_{5ij}) + \Lambda_{17}(X_{2ij}X_{4ij}X_{5ij}) + \Lambda_{18}(X_{3ij}X_{4ij}X_{5ij}) + r_{ij}$$

STAGE 2:

Level 1:

$$ITBS\_RES\_R\_96_{ij} = \beta_{0j} + \beta_{1j} ITBS\_RES\_R\_95_{ij} + \beta_{2j} ITBS\_RES\_M\_95_{ij} + \delta_{ij}$$

Level 2:

$$\beta_{0j}\ \gamma_{00} + \gamma_{01} W_{1j} + \gamma_{02} W_{2j} + \ldots + \gamma_{010} W_{10j} + u_{0j}$$

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1} W_{1j} + \gamma_{k2} W_{2j} + \ldots + \gamma_{k10} W_{10j}$$

for $k = 1, 2$.

$E[u_{0j}] = 0$, $\text{Var}[u_{0j}] = \tau^2$, and $u_{0j} \perp \delta_{ij}$

$$SEI_j^* = u_{0j}^*$$

Model 8

Level 1:

$$ITBS\_R\_96_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \beta_{3j} X_{3ij} + \beta_{4j} X_{4ij} + \beta_{5j} X_{5ij} + \beta_{6j} X_{6ij} + \beta_{7j} X_{7ij} + \beta_{8j} X_{8ij} + \beta_{9j}(X_{1ij}X_{4ij}) + \beta_{10j}(X_{2ij}X_{4ij}) + \beta_{11j}(X_{3ij}X_{4ij}) + \beta_{12j}(X_{1ij}X_{5ij}) + \beta_{13j}(X_{2ij}X_{5ij}) + \beta_{14j}(X_{3ij}X_{5ij}) + \beta_{15j}(X_{4ij}X_{5ij}) + \beta_{16j}(X_{1ij}X_{4ij}X_{5ij}) + \beta_{17j}(X_{2ij}X_{4ij}X_{5ij}) + \beta_{18j}(X_{3ij}X_{4ij}X_{5ij}) + \beta_{19j} ITBS\_R\_95_{ij} + \beta_{20j} ITBS\_M\_95_{ij} + \delta_{ij}$$

where

$$\delta_{ij} \overset{iid}{\sim} N(0,\sigma^2).$$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_{1j} + \gamma_{02} W_{2j} + \ldots + \gamma_{010} W_{10j} + u_{0j}$$

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1} W_{1j} + \gamma_{k2} W_{2j} + \ldots + \gamma_{k10} W_{10j}$$

for $k = 1, 2, \ldots, 20$.

$E[u_{0j}] = 0$, $Var[u_{0j}] = \tau^2$, and $u_{0j} \perp \delta_{ij}$

$$SEI_j^* = u_{0j}^*$$

Models 9, 10, 11, and 12 utilize three years of data to predict a fourth. They were designed to compare the results of these analyses with the results of Models 1 through 8 that use only one year of prediction in conjunction with a wealth of contextual variables. Models 9 and 10 do not utilize any contextual variables but rather depend on individual student growth histories to account for the variance normally associated with contextual variables. Models 11 and 12 add contextual variables to the equations, Model 11 at the conditioning level and Model 12 at both the student and conditioning levels. Appropriate equations for Models 9 through 12 follow:

Model 9

$$ITBS\_R\_96_{ij} = \Lambda_0 + \Lambda_1 ITBS\_R\_95_{ij} + \Lambda_2 ITBS\_M\_95_{ij} + \Lambda_3 ITBS\_R\_94_{ij} + \Lambda_4 ITBS\_M\_94_{ij} + \Lambda_5 ITBS\_R\_93_{ij} + \Lambda_6 ITBS\_M\_93_{ij} + \varepsilon_{ij}$$

$$SEI_j = \frac{\sum_{i=1}^{N_j} \varepsilon_{ij}}{N_j}$$

Model 10

Level 1:

$$ITBS\_R\_96_{ij} = \beta_0 + \beta_1 ITBS\_R\_95_{ij} + \beta_2 ITBS\_M\_95_{ij} + \beta_3 ITBS\_R\_94_{ij} + \beta_4 ITBS\_M\_94_{ij} + \beta_5 ITBS\_R\_93_{ij} + \beta_6 ITBS\_M\_93_{ij} + \delta_{ij}$$

Level 2:

$\beta_{0j} = \gamma_{00} + u_{0j}$
$\beta_{kj} = \gamma_{k0}$

for $k = 1, 2, ..., 6$.

$E[u_{0j}] = 0$, $Var[u_{0j}] = \tau^2$, and $u_{0j} \perp \delta_{ij}$

$$SEI_j^* = u_{0j}^*$$

Model 11

Level 1:

$$\begin{aligned} ITBS\_R\_96_{ij} = {} & \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij} + \beta_{5j}X_{5ij} + \beta_{6j}X_{6ij} + \beta_{7j}X_{7ij} + \beta_{8j}X_{8ij} + \\ & \beta_{9j}(X_{1ij}X_{4ij}) + \beta_{10j}(X_{2ij}X_{4ij}) + \beta_{11j}(X_{3ij}X_{4ij}) + \beta_{12j}(X_{1ij}X_{5ij}) + \beta_{13j}(X_{2ij}X_{5ij}) + \\ & \beta_{14j}(X_{3ij}X_{5ij}) + \beta_{15j}(X_{4ij}X_{5ij}) + \beta_{16j}(X_{1ij}X_{4ij}X_{5ij}) + \beta_{17j}(X_{2ij}X_{4ij}X_{5ij}) + \\ & \beta_{18j}(X_{3ij}X_{4ij}X_{5ij}) + \beta_{19j}ITBS\_R\_95_{ij} + \beta_{20j}ITBS\_M\_95_{ij} + \beta_{21}ITBS\_R\_94_{ij} + \\ & \beta_{22}ITBS\_M\_94_{ij} + \beta_{23}ITBS\_R\_93_{ij} + \beta_{24}ITBS\_M\_93_{ij} + \delta_{ij} \end{aligned}$$

where

$\delta_{ij} \overset{iid}{\sim} N(0,\sigma^2)$.

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_{1j} + \gamma_{02}W_{2j} + \ldots + \gamma_{010}W_{10j} + u_{0j}$$
$$\beta_{kj} = \gamma_{k0} + \gamma_{k1}W_{1j} + \gamma_{k2}W_{2j} + \ldots + \gamma_{k10}W_{10j}$$

for $k = 1, 2, ..., 24$.

$E[u_{0j}] = 0$, $Var[u_{0j}] = \tau^2$, and $u_{0j} \perp \delta_{ij}$

$$SEI_j^* = u_{0j}^*$$

Model 12

STAGE 1:

$Y_{ij} = \Lambda_0 + \Lambda_1 X_{1ij} + \Lambda_2 X_{2ij} + \Lambda_3 X_{3ij} + \Lambda_4 X_{4ij} + \Lambda_5 X_{5ij} + \Lambda_6 X_{6ij} + \Lambda_7 X_{7ij} + \Lambda_8 X_{8ij} + \Lambda_9(X_{1ij}X_{4ij}) + \Lambda_{10}(X_{2ij}X_{4ij}) + \Lambda_{11}(X_{3ij}X_{4ij}) + \Lambda_{12}(X_{1ij}X_{5ij}) + \Lambda_{13}(X_{2ij}X_{5ij}) + \Lambda_{14}(X_{3ij}X_{5ij}) + \Lambda_{15}(X_{4ij}X_{5ij}) + \Lambda_{16}(X_{1ij}X_{4ij}X_{5ij}) + \Lambda_{17}(X_{2ij}X_{4ij}X_{5ij}) + \Lambda_{18}(X_{3ij}X_{4ij}X_{5ij}) + r_{ij}$

where Y_{ij} is ITBS_R_96$_{ij}$, ITBS_R_95$_{ij}$, ITBS_M_95$_{ij}$, ITBS_R_94$_{ij}$, ITBS_M_94$_{ij}$, ITBS_R_93$_{ij}$, ITBS_M_93$_{ij}$, ITBS_R_92$_{ij}$, and ITBS_M_92$_{ij}$. These will produce ITBS_RES_R_96$_{ij}$, ITBS_RES_R_95$_{ij}$, ITBS_RES_M_95$_{ij}$, ITBS_RES_R_94$_{ij}$, ITBS_RES_M_94$_{ij}$, ITBS_RES_R_93$_{ij}$ and ITBS_RES_M_93$_{ij}$, respectively.

STAGE 2:

Level 1:

$\text{ITBS\_RES\_R\_96}_{ij} = \beta_{0j} + \beta_{1j}\text{ITBS\_RES\_R\_95}_{ij} + \beta_{2j}\text{ITBS\_RES\_M\_95}_{ij} + \beta_{3j}\text{ITBS\_RES\_R\_94}_{ij} + \beta_{4j}\text{ITBS\_RES\_M\_94}_{ij} + \beta_{5j}\text{ITBS\_RES\_R\_93}_{ij} + \beta_{6j}\text{ITBS\_RES\_M\_93}_{ij} + \delta_{ij}$

where

$\delta_{ij} \overset{iid}{\sim} N(0,\sigma^2)$.

Level 2:

$\beta_{0j} = \gamma_{00} + \gamma_{01}W_{1j} + \gamma_{02}W_{2j} + \ldots + \gamma_{010}W_{10j} + u_{0j}$

$\beta_{kj} = \gamma_{k0} + \gamma_{k1}W_{1j} + \gamma_{k2}W_{2j} + \ldots + \gamma_{k10}W_{10j}$

for $k = 1, 2, ..., 6$.

$E[u_{0j}] = 0$, $Var[u_{0j}] = \tau^2$, and $u_{0j} \perp \delta_{ij}$

$$SEI_j^* = u_{0j}^*$$

Model 13 is a three level HLM gain model similar to the model proposed by Bryk and Thum (1996). It is compared to Models 14 and 15, models that are comparable to Model 13 except that they are status models, not gain score models. (A status model is a model that uses actual test scores or residuals of actual test scores rather than gain scores as the basic unit of analysis. All Models in this paper except Model 13 are status models.) Appropriate equations for Models 13, 14, and 15 follows:

Model 13

Level 1:

$\text{ITBS\_GAIN\_R95\_R96}_{ijk} = \pi_{0jk} + \pi_{1jk}\text{ITBS\_R\_95}_{ijk} + \pi_{2jk}\text{ITBS\_M\_95}_{ijk} + \pi_{3jk}\text{ITBS\_R\_94}_{ijk} + \pi_{4jk}\text{ITBS\_M\_94}_{ijk} + \pi_{5jk}\text{ITBS\_R\_93}_{ijk} + \pi_{6jk}\text{ITBS\_M\_93}_{ijk} + \varepsilon_{ijk}$

where $\varepsilon_{ijk} \overset{iid}{\sim} N(0, 1)$ and *i*, *j* both refer to the same student in school *k*.

Level 2:

$\pi_{pjk} = \beta_{p0k} + \beta_{p1k}\text{BLACK}_{jk} + \beta_{p2k}\text{HISPANIC}_{jk} + \beta_{p3k}\text{GENDER}_{jk} + r_{pjk}$

where $r_{pjk} \overset{iid}{\sim} N(0,T)$ and $r_{pjk} \perp \varepsilon_{ijk}$.

Level 3:

$\beta_{p0k} = \gamma_{00k} + u_{p0k}$

$\beta_{pqk} = \gamma_{p0k}$ for $q = 1, 2, 3$

$E[u_{p0k}] = 0$, $Var[u_{p0k}] = \Delta^2$, $u_{p0k} \perp r_{pjk}$ and $u_{p0k} \perp \varepsilon_{ijk}$.

$$SEI_k^* = u_{00k}^*$$

Model 14

STAGE 1:

$Y_{ij} = \Lambda_0 + \Lambda_1\text{BLACK}_{ij} + \Lambda_2\text{HISPANIC}_{ij} + \Lambda_3\text{GEND R}_{ij} + \varepsilon_{ij}$

where Y_{ij} is ITBS_R_96$_{ij}$, ITBS_R_95$_{ij}$, ITBS_M_95$_{ij}$, ITBS_R_94$_{ij}$, ITBS_M_94$_{ij}$, ITBS_R_93$_{ij}$ and ITBS_M_93$_{ij}$..

STAGE 2:

Level 1:

$\text{ITBS\_RES\_R\_96}_{ij} = \beta_{0j} + \beta_{1j}\text{ITBS\_RES\_R\_95}_{ij} + \beta_{2j}\text{ITBS\_RES\_M\_95}_{ij} + \beta_{3j}\text{ITBS\_R S\_R\_94}_{ij} + \beta_{4j}\text{ITBS\_RES\_M\_94}_{ij}$

$+ \beta_{5j}ITBS\_RES\_R\_93_{ij} +$
$\beta_{6j}ITBS\_RES\_M\_93_{ij} + \delta_{ij}$

Level 2:
$\beta_{0j} = \gamma_{00} + u_{0j}$
$\beta_{kj} = \gamma_{k0}$ for $k = 1, 2, ..., 6$.

where $E[u_{ij}] = 0$, $Var[u_{0j}] = \tau^2$, and $u_{0j} \perp \delta_{ij}$

$$SEI_k^* = u_{00k}^*$$

Model 15

Level 1:

$ITBS\_R\_96_{ij} = \beta_{0j} + \beta_{1j}BLACK_{ij} +$
$\beta_{2j}HISPANIC_{ij} + \beta_{3j}GENDER_{ij} +$
$\beta_{4j}ITBS\_R\_95_{ij} + \beta_{5j}ITBS\_M\_95_{ij} +$
$\beta_{6j}ITBS\_R\_94_{ij} + \beta_{7j}ITBS\_M\_94_{ij} +$
$\beta_{8j}ITBS\_R\_93_{ij} + \beta_{9j}ITBS\_M\_93_{ij} + \delta_{ij}$

where
$\delta_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

Level 2:

$\beta_{0j} = \gamma_{00} + u_{0j}$
$\beta_{kj} = \gamma_{k0}$ for $k = 1, 2, ..., 9$.

$E[u_{0j}] = 0$, $Var[u_{0j}] = \tau^2$, and $u_{0j} \perp \delta_{ij}$

$$SEI_j^* = u_{0j}^*$$

The results produced by the OLS regression models (Models 1, 3, 9) were adjusted for shrinkage by the following procedure:

For District:

$$\mu = \frac{\sum_{j=1}^{J}\sum_{i=1}^{N_j}\varepsilon_{ij}}{\sum_{j=1}^{J}N_j}$$

$$\sigma^2 = \frac{\sum_{j=1}^{J}\sum_{i=1}^{N_i}(\varepsilon_{ij} - \mu)^2}{\sum_{j=1}^{J}N_j}$$

For each school j:

$$\mu_j = \frac{\sum_{i=1}^{N_j}\varepsilon_{ij}}{N_j}$$

$$\sigma_j^2 = \frac{\sum_{i=1}^{N_i}(\varepsilon_{ij} - \mu_j)^2}{N_j}$$

The shrinkage coefficient is,

$$\lambda_j = \frac{\sigma^2}{\sigma^2 + \frac{\sigma_j^2}{N_j}}$$

then the shrinkage adjusted SEI is

$$SEI^* = \lambda_j\mu_j + (1-\lambda_j)\mu$$

SEI's produced by HLM are already adjusted for shrinkage.

Teacher Effect

Seventeen different OLS regression and HLM models were investigated to determine their reliability and appropriateness for measuring teacher effect. Figure 2 contains descriptions of these models. The first twelve models use the same equations to generate the residuals that were used in the school level models. The results are then adjusted for shrinkage through the use of the following formulas:

CEIs

Models 1 - 12

$CEI_{mj} = m^{th}$ classroom in school j.

CEI_{mj} is obtained by aggregating the student residuals by classroom

The shrinkage adjustment is as follows:

for the residuals ε_{imj} or δ_{imj} for i^{th} student in classroom *m* in school *j*, calculated with respect to the district

$$\nu = \frac{\sum_{j=1}^{J}\sum_{m=1}^{M_j}\sum_{i=1}^{N_{mj}} \varepsilon_{imj}}{\sum_{j=1}^{J}\sum_{m=1}^{M_j} N_{mj}}$$

$$\tau^2 = \frac{\sum_{j=1}^{J}\sum_{m=1}^{M_j}\sum_{i=1}^{N_{mj}} (\varepsilon_{imj} - \nu)^2}{\sum_{j=1}^{J}\sum_{m=1}^{M_j} N_{mj}}$$

$$\nu_{mj} = \frac{\sum_{i=1}^{N_{mj}} \varepsilon_{imj}}{N_{mj}}$$

Insert Figure 2 Here

Insert Figure 2 (cont.)

$$\tau_{mj}^2 = \frac{\sum_{i=1}^{N_{mj}} (\varepsilon_{imj} - \nu_{mj})^2}{N_{mj}}$$

The shrinkage coefficient is

$$\lambda_{mj} = \frac{\tau^2}{\tau^2 + \frac{\tau_{mj}^2}{N_{mj}}}$$

Hence, the shrinkage adjusted CEIs for models 1 to 12 are

$$CEI^*_{mj} = \lambda_{mj}\nu_{mj} + (1-\lambda_{mj})\,\nu$$

Models 13, 14, and 15 are two-level HLM models with classroom as the conditioning level instead of school. These models produce empirical Bayes estimates around the District mean and thus produce systemwide teacher effectiveness indices. The results of these models can be directly compared to the results of Models 1-12. Model 13 is a one-stage, two-level HLM while Models 14 and 15 are two-stage, two-level models. Model 14 assumes fixed slopes while Model 15 assumes random slopes.

Model 13

Level 1:

$$ITBS\_R\_96_{ij} = \beta_{0j} + \beta_{1j}ITBS\_R\_95_{ij} + \beta_{2j}ITBS\_M\_95_{ij} + \delta_{ij}$$

where $\delta_{ij} \overset{iid}{\sim} N(0,\sigma^2)$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_{1j} + \gamma_{02}T_{2j} + \ldots + \gamma_{010}T_{10j} + u_{0j}$$

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1}T_{1j} + \gamma_{k2}T_{2j} + \ldots + \gamma_{k10}T_{10j}$$

for $k = 1, 2$.

$E[u_{0j}] = 0$, $Var[u_{0j}] = \tau^2$, and $u_{0j} \perp \delta_{ij}$

$$CEI_j^* = u_{0j}^*$$

Model 14

Level 1:

$$ITBS\_RES\_R\_96_{ij} = \beta_{0j} + \beta_{1j}ITBS\_RES\_R\_95_{ij} + \beta_{2j}ITBS\_RES\_M\_95_{ij} + \delta_{ij}$$

where $\delta_{ij} \overset{iid}{\sim} N(0,\sigma^2)$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_{1j} + \gamma_{02}T_{2j} + \ldots + \gamma_{010}T_{10j} + u_{0j}$$

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1}T_{1j} + \gamma_{k2}T_{2j} + \ldots + \gamma_{k10}T_{10j}$$

for $k = 1, 2$.

$E[u_{0j}] = 0$, $Var[u_{0j}] = \tau^2$, and $u_{0j} \perp \delta_{ij}$

$$CEI_j^* = u_{0j}^*$$

Model 15

Level 1:

$$ITBS\_RES\_R\_96_{ij} = \beta_{0j} + \beta_{1j}ITBS\_RES\_R\_95_{ij} + \beta_{2j}ITBS\_RES\_M\_95_{ij} + \delta_{ij}$$

where $\delta_{ij} \overset{iid}{\sim} N(0,\sigma^2)$

Level 2:

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1}T_{1j} + \gamma_{k2}T_{2j} + \ldots + \gamma_{k10}T_{10j} + u_{kj}$$

for $k = 0, 1, 2$.

$E[u_{0j}] = 0$, $Var[u_{0j}] = \tau^2$, and $u_{0j} \perp \delta_{ij}$

$$CEI_j^* = u_{0j}^*$$

Model 16 is a three-level HLM model that produces empirical Bayes estimates around the school mean for each teacher. The results produced by this model are compared to Model 17. Model 17 is identical to Model 7 except that the teacher level residuals are calculated about the school means rather than about the district mean. This should enable a direct comparison with the results produced by Model 16. Appropriate equations follow:

Model 16

Level 1:

$$ITBS\_R\_96_{ijk} = \pi_{0jk} + \pi_{1jk}ITBS\_R\_95_{ijk} + \pi_{2jk}ITBS\_M\_95_{ijk} + \varepsilon_{ijk}$$

where $\varepsilon_{ijk} \overset{iid}{\sim} N(0,\sigma^2)$.

Level 2:

$$\pi_{pjk} = \beta_{p0k} + \sum_{q=1}^{10} \beta_{pqk}T_{qjk} + \delta_{pjk}$$

where $\delta_{pjk} \overset{iid}{\sim} N(0,T)$ and $\delta_{pjk} \perp \varepsilon_{ijk}$.

Level 3:

$$\beta_{00k} = \gamma_{000} + u_{00k}$$

$$\beta_{pqk} = \gamma_{pq0}$$ for all other p and q.

$$CEI_{jk}^* = \gamma_{0jk}^*$$

Model 17

STAGE 1:

$$Y_{ij} = \Lambda_0 + \Lambda_1X_{1ij} + \Lambda_2X_{2ij} + \Lambda_3X_{3ij} + \Lambda_4X_{4ij} + \Lambda_5X_{5ij} + \Lambda_6X_{6ij} + \Lambda_7X_{7ij} + \Lambda_8X_{8ij} + \Lambda_9(X_{1ij}X_{4ij}) + \Lambda_{10}(X_{2ij}X_{4ij}) + \Lambda_{11}(X_{3ij}X_{4ij}) + \Lambda_{12}(X_{1ij}X_{5ij}) + \Lambda_{13}(X_{2ij}X_{5ij}) + \Lambda_{14}(X_{3ij}X_{5ij}) + \Lambda_{15}(X_{4ij}X_{5ij}) + \Lambda_{16}(X_{1ij}X_{4ij}X_{5ij}) + \Lambda_{17}(X_{2ij}X_{4ij}X_{5ij}) + \Lambda_{18}(X_{3ij}X_{4ij}X_{5ij}) + r_{ij}$$

STAGE 2:

Level 1:

$$ITBS\_RES\_R\_96_{ij} = \beta_{0j} + \beta_{1j}ITBS\_RES\_R\_95_{ij} + \beta_{2j}ITBS\_RES\_M\_95_{ij} + \delta_{ij}$$

Level 2:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{kj} = \gamma_{k0}$$ for $k = 1, 2$.

$E[u_{0j}] = 0$, $Var[u_{0j}] = \tau^2$, and $u_{0j} \perp \delta_{ij}$

The student residuals, δ_{ij}s, are calculated with respect to each school and shrinkage adjusted to obtain CEI_j^*.

Results

School Effectiveness Indices

The most efficient way to discuss results is to present all data and then discuss all results simultaneously. With that end in mind, the following tables are presented:

Table 1 Correlations Between and Among The School Effectiveness Indices Produced By Each of the Models, Grade 6

Table 2 Correlations Between and Among The School Effectiveness Indices Produced By Each of the Models, Grade 8

As mentioned previously the major difference between the grade six and grade eight samples is that the sixth grade represents 127 relatively homogeneous schools while the eighth grade consists of only 26 relatively heterogeneous schools. Put another way, there is far more within school variance relative to between school variance at the eighth grade level than there is at the sixth grade level. The eighth grade was included in this study to insure that results were not situation specific, i.e., did not only apply to situations where there were large numbers of relatively homogeneous schools.

The reader will recall that, at the school level, we are investigating six questions. First, is there any practical difference between effectiveness indices produced by two-stage versus one-stage models? Second, is there any difference between effectiveness indices produced by HLM models assuming fixed versus random slopes? Third, does a three-level HLM model that uses student gain scores as the outcome variable produce results that are similar to those produced by status-based models? Fourth, how free from bias relative to important student and school level contextual variables and pretest scores are the various models? Fifth, can a longitudinal student growth curve approach to predicting school effect produce bias free results? Finally, although not explicitly stated, is there a best model for estimating school effect?

In examining the School Effectiveness Indices one-stage versus two-stage models, one generally finds little difference between the two. Correlations, between the products of Models 1 and 3 (OLS Regression) were .9595 at grade 6 and .9403 at grade 8; between Models 2 and 5 (HLM-no school level variables) were .9545 and .9415, respectively; and between Models 7 and 8 were .9153 and .5306. The relatively low correlations between Models 7 and 8 were primarily due to the fact that no three-way interactions, no math predictor, and no census data could be included in the one-stage eighth grade HLM model. In addition, the correlations of residuals produced by the one-stage HLM models with student level contextual variables suggest that HLM one-stage models carry suppresser effects that are not found in OLS regression models or two-stage HLM models. When this occurrence is coupled with the inability to include important school level contextual variable in the one-stage HLM models, resulting in unsatisfactory correlations between the results produced by the one-stage HLM full model and those important school level contextual variables, it is concluded that two-stage HLM models are more appropriate for use in estimating school effect.

In investigating the fixed versus random slopes issues, School Effectiveness Indices produced by the two types of models were highly correlated when working with a large number of schools (grade 6 correlations between Models 4 and 5 and Models 6 and 7 were .9810 and .9867, respectively) and moderately correlated when working with a smaller number of schools at grade 8 (.9377 and .8126, respectively). These comparisons were all computed with two-stage models, since one-stage HLM full models assuming random slopes could not be solved. These models produced low correlations with student level variable and, when school level conditioning variables were added, zero correlations with school level variables. The authors believe that the differences at grade 8 occurred because the fixed models do not account for the larger variation present in the slopes of a small number of schools. Given these slight differences, the authors suggest the use of random models in estimating school effect.

With regard to the issue of the gain score model with limited conditioning variables producing results similar to those produced by similar status-based models, there are two answers. An earlier paper by Weerasinghe, et. al, (1997) arrived at the conclusion that, if the same predictor variables are used in the two models, the results are very similar. This conclusion is supported by the relatively high correlations between the School Effectiveness Indices produced by Model 13 and Model 14 (.9535). However, Weerasinghe, et.al., (1997) found that two level HLM status-models are far more convenient, efficient, and less fragile than the three level gain model. In the two-level models, far more Level 1 and Level 2 variables can be introduced to obtain complex models without any biases to the conditioning variables. The three level model is also very sensitive to multicolinearity and low variances in conditioning variables.

Returning to this analysis, it is clear that the School Effectiveness Indices produced by Models 13 and 14 are different from those produced by other models utilized in this study. Much, but not all of this difference is due to the lack of conditioning variables included in Models 13 and 14. Correlations of results produced by these models with important school contextual variable are sufficiently high as to suggest a major bias in the indices produced. This finding demands that one either add additional school level conditioning variable to these models, or failing that, go to less complex models that will allow more conditioning variables. The remaining difference is due to missing data deriving from the use of three years of student score for prediction versus one year of student score in concert with a rich array of contextual information. Since the authors are charged with the responsibility of determining school effect over a one year period, we believe that the one year approach maximizes available information and is more appropriate to the task.

Most of the measures produced by the various models are free from significant bias at the student level. Bias enters in at the school level unless important contextual variables are included as conditioning variables in an HLM model. None of the indices produced by the various models correlate significantly with pretest scores.

With regard to longitudinal models, it is clear that longitudinal models produce results that are very similar to one-year models with identical conditioning variables (Models 8 vs. 11, .9626 grade 6, .9580 grade 8; Models 7 vs. 12, .9547 grade 6, .9162 grade 8). These small differences can easily be attributed to missing data that occurs in the longitudinal analyses. It is also clear that without the inclusion of school level conditioning variables, longitudinal models produce results that carry severe biases against schools serving minority and poor students. These biases are far more pronounced than even the OLS regression models and HLM models that utilized one year of prediction and did not control for school level contextual variables (Models 1 through 5). It is also interesting to note that the correlation between Models 10A and 10, one with three years of prediction, the other with four is .9992. Thus the additional year provides no additional information and costs about 5% of the population.

Conclusions on SEI

Based on the analyses conducted through this study, the authors believe that HLM two-stage, two-level, random models with a full range of student and school level contextual variables produce the most bias free estimates of school effect. The model of choice is Model 6.

Teacher Effectiveness Indices

The following Tables present results relative to the teacher effectiveness indices:

Table 7 Correlations Between and Among The Teacher Effectiveness Indices Produced By Each of the Models, Grade 6

Table 8 Correlations Between and Among The Teacher Effectiveness Indices Produced By Each of the Models, Grade 8

Table 9 Correlations of The Teacher Effectiveness Indices with Important Teacher Contextual Variables, Grade 6

Table 10 Correlations of The Teacher Effectiveness Indices with Important Teacher Contextual Variables, Grade 8

Table 11 Correlations of The Teacher Effectiveness Indices with Important Student Contextual Variables, Grade 6

Table 12 Correlations of The Teacher Effectiveness Indices with Important Student Contextual Variables, Grade 8

Note that results for Model 16 are not included in any of the teacher tables. Model 16 (three-level HLM, random slopes at level 2, fixed slopes at level 3) was designed to allow the inclusion of classroom level conditioning variables at level 2. It was calculated in the form specified by Model 16 and in every other conceivable combination including two-stage models. These models would not run with a full array of conditioning variables at the teacher and school levels. The best we could do was enter four conditioning variables at each level. The computed effectiveness indices were dependent upon the conditioning variables included in the equations. Since all conditioning variables are included in the equations for specific reasons, it is repugnant not to use all available relevant information and thus three-level models proved too fragile to run and had to be abandoned.

In examining the other models, note first that the correlations between the various combinations of models (Tables 7 and 8) show little difference among the first eight models. One-stage OLS regression (Model 1) and one-stage HLM (Model 8) are the only models that differ slightly and systematically from the others with correlations ranging from .9363 to .9919 at grade 6 and .8543 to .9680 at grade 8. The remaining model intercorrelations range from .9506 to .9999 at grade 6 and .9317 to .9999 at grade 8. In particular, the two stage HLM models, 4 through 7, have intercorrelations at or above .9997. (The last is not particularly surprising since the models are computed from extremely closely related sets of student residuals.)

The longitudinal models, 9 and 12, show mostly moderate intercorrelations with the other longitudinal models and themselves at grade 6 (.8709 to .9396) and grade 8 (.8421 to .9132) while the longitudinal one-stage HLM models show higher intercorrelations at both grades (.9929 to .9993 at grade 6 and .9427 to .9853 at grade 8). In general, the correlations of the longitudinal models with the other models are lower at both grades (generally about .8800 at grade 6 and .8300 at grade 8 with several exceptions that are somewhat higher.) The two-level student-teacher HLM models, Models 13, 14 and 15, show high intercorrelations at both grade 6 and 8 (>.90). Nothing correlates very highly with Model 17.

The discussion of the intercorrelations of the teacher indices models is intentionally terse, because the important information about these models comes from the examination of their relationship to the classroom level conditioning variables in Tables 9 and 10. All of the models, with the exception of Models 13, 14, and 15, show unacceptably high correlations with SES variables at the classroom level (free lunch and the census variables). Correlations with free lunch at grade 6 range from -.1073 to -.3153 and correlations with census income at grade 8 range from .1710 to .4314. In plain words, with the exception of Models 13, 14, 15, all of the models are biased against classrooms with higher percentages of low SES students. Where the classroom level conditioning variables are included in the second stage of a two level HLM model, all intercorrelations disappear.

The degree of bias in the other models varies. The one-stage OLS model (Model 1) is the least biased at grade 6 while the one stage fixed slopes HLM model and the longitudinal two-level HLM with fixed slopes (Models 8, 10 and 10A) are the most biased at grade 6. At grade 8, longitudinal Model 12 is the least biased and Models 8, 10 and 10A are the most biased. Of the least biased models, the OLS model at grade 6 comes close to being acceptable as a usable model without the addition of classroom variables.

Conclusions on TEI

Now, considering the questions posed at the beginning of the paper, the responses are immediate. All models estimating classroom effects are biased unless classroom level variables are included as conditioning variables. Thus questions of OLS versus HLM, one-stage versus two-stage, fixed versus random, and one-year versus longitudinal all are insignificant without the elimination of bias in classroom level SES-related variables. Models 13, 14, and 15, all two-level student-teacher HLM models, produce acceptable results. However, because one-stage HLM models often carry suppressor effects and fixed models do not account for large variations in teacher slopes, it is recommended that a two-stage, two-level random model be employed with a full range of student and classroom level contextual variables. Thus, the model of choice is Model 15.

Discussion

The information in these investigations has brought several issues into sharp focus for the authors. The original foray into identifying effective schools conducted in Dallas in the 1980s (Webster and Olson, 1988) resulted in a method that was fair at the student level, but less so at the school level. The current set of research studies begun in the early 1990s (Mendro and Webster, 1993; Webster, Mendro, and Almaguer, 1994) solved the problems identified at the school level first through an OLS model that included interactions among the student level variables and then refined the model with the HLM model including school variables explicitly at the second level.

In designing this study, the authors had the naive expectation that they would be able to complete a set of analyses that would give us a set of answers to guide future analyses and efforts in our own attempts to determine effective schools and teachers and that extensive future research of this type would not be necessary. We were wrong. The results at 8th grade which show unexpected problems with models containing few level two data points (number of schools) and the results for the teacher indices which showed the correlations with classroom variables indicate that further research on both fronts will have to continue for the foreseeable future.

Also, the authors had once speculated, given the similarities among our previous sets of results, that any carefully thought out regression approach, OLS or HLM would produce acceptable results (Webster et. al. 1995). The cumulative effect of our prior research and these studies now indicates that our speculation was premature and probably wrong as well.

It is becoming clear to us that no assertions about models and their efficacy can be taken at face value without extensive trials of the models and careful comparisons of their output. We suspect that there may be ways to adapt OLS models to include second level conditioning variables (for Teacher or School Effectiveness Indices) that may produce more acceptable results than a number of the models tested here. Further, for our own effectiveness programs in our District, we need to carefully compare teacher models that employ classroom-level conditioning variables. However, the critical point is that we are no longer willing to make assumptions about models without careful examinations of the practical results.

This does not say that our research has failed to result in some general conclusions about models for identifying effective teachers and schools. Until we arrive at a model with better characteristics, we note that school models that are two-stage HLM models, that eliminate student level characteristics at the first stage and employ relevant conditioning variables at the second stage with random effects present the best choice for a school effects model. The choice of one-versus two-stage is clear because of suppresser effects.

The school conditioning variables are necessary to control these variables. Finally, the choice of a random model seems to offer the best method of controlling variables which have not been explicitly controlled in the model.

For teacher models, clearly this study has shown the need to control classroom level conditioning variables. Future models will have to take that as a given element or will have to show that they do so intrinsically to be seriously considered as acceptable models. At this point, however, the authors intend to apply a two-stage, two-level student-classroom HLM model that eliminates student level characteristics at the first stage and employs relevant conditioning variables at the second stage with random effects to estimate teacher effect.

References

Aiken, L.S. and West, S.G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. Newburg Park: Sage

Bano, S.M. (1985). *The Logic of Teacher Incentives*. Washington, D.C.: National Association of State Boards of Education.

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newburg Pass, California: Sage Publications

Bryk, A.S., Raudenbush, S.W., Seltzer, M., and Congdon, R. (1988). *An Introduction to HLM: Computer Program User's Guide* (2nd ed.) Chicago, Ill: University of Chicago

Bryk, S.A. and Thum, Y.M., (1996). *Assessing the Productivity of Chicago Schools under Reform.* 1996 NORC Professional Development Training Session on Data Analysis: Concepts and Applications, Chicago

Dempster, A.P., Rubin, D.B., and Tsutakawa, R.V., (1981). *Estimation in Covariance Components Models*. Journal of the American Statistical Association, 76, 341-353.

Elston, R.C. and Grizzle, J.E. (1962). *Estimation of Time Response Curves and Their Confidence Bands*. Biometrics, 18, 148-159.

Felter, M. and Carlson, D. (1985). Identification of Exemplary Schools on a Large Scale. In Austin and Gerber (eds.), *Research on Exemplary Schools*. New York: Academic Press, 83-96

Goldstein, H., (1987). *Multilevel Models in Educational and Social Research*, New York: Oxford University Press.

Henderson, C.R., (1984). *Applications of Linear Models in Animal Breeding*. Guelph, Canada: University of Guelph.

Kirst, M. (1986). New Directions for State Education Data Systems. *Education and Urban Society*, 18, 2, 343-357.

Klitgaard, R.E. and Hall, G.R. (1973). *A Statistical Search for Unusually Effective Schools*. Santa Monica, Ca.: Rand Corporation.

Laird, N.M. and Ware, H. (1982). *Random-Effects Models for Longitudinal Data*, Biometrics, 38-963-974.

McKenzie, D. (1983). School Effectiveness Research: A Synthesis and Assessment. In P. Duttweiler (ed.), *Educational Productivity and School Effectiveness*. Austin, Texas: Southwest Educational Development Laboratory.

Mason, W.M., Wong, G.Y. and Entwistle, B. (1984). Contextual Analysis Through the Multilevel Linear Model. In Leinhardt (ed.) *Sociological Methodology*, 1983-84, San Francisco: Josey-Bass, pp. 72-103.

Mendro, R.L. and Webster, W.J. (1993). *Using School Effectiveness Indices to Identify and Reward Effective Schools.* Paper presented at the Rocky Mountain Research Association, October, 1993, Las Cruces, New Mexico.

Mendro, R. L., Webster, W. J., Bembry, K. L., and Orsak, T. H. (1995). *An Application of Hierarchical Linear Modeling in Determining School Effectiveness.* A paper presented at the 1995 Annual Meeting of the American Educational Research Association, San Francisco, CA.

Millman, J. (ed.) (1981). *Handbook of Teacher Evaluation*. Beverly Hills, California, Sage.

Rosenburg, B. (1973). *Linear Regression With Randomly Dispersed Parameters*. Biometrika, 60, 61-75.

Saka, T. (1984). *Indicators of School Effectiveness: Which are the Most Valid and What Impacts Upon Them*? Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA: (ERIC No. ED 306277)

Thum, Y. and Bryk, A. (1997). Value-Added Productivity Indicators: The Dallas System in *Assuring Accountability . . . ? Using Gains in Student Learning to Evaluate Teachers and Schools*, (Ed. by J. Millman) Newbury Park, CA, Sage Publications. in press.

Webster, W. J. and Olson, G. H. (1988). A Quantitative Procedure for the Identification of Effective Schools. *Journal of Experimental Education*, 56, 213-219.

Webster, W.J., Mendro, R.L., and Almaguer, T. (1994). Effectiveness Indices: A "Value Added" Approach to Measuring School Effect. *Studies in Educational Evaluation*. 20, 113-145

Webster, W. J., Mendro, R. L., Bembry, K. L., and Orsak, T. H. (1995). *Alternative Methodologies for Identifying Effective Schools*. A paper presented at the 1995 Annual Meeting of the American Educational Research Association, San Francisco, CA.

Webster, W. J. and Mendro, R. L. (1997). *The Dallas Value-Added Accountability System in Assuring Accountability . . . ? Using Gains in Student Learning to Evaluate Teachers and Schools*, (Ed. by J. Millman) Newbury Park, CA, Sage Publications. in press.

Webster, W. J., Mendro, R. L., Orsak, T. H., and Weerasinghe, D. (1996). *The Applicability of Selected Regression and Hierarchical Linear Models To The Estimation of School and Teacher Effects.*. A paper presented at the Annual Meeting of the American Educational Research Association, March, 1996, New York, NY.

Webster, W. J., Mendro, R. L., Orsak, T., Weerasinghe, D., and Bembry, K. (1997). Little Practical Difference and Pie in the Sky: A Response to Thum and Bryk and a Rejoinder to Sykes in *Assuring Accountability . . . ? Using Gains in Student Learning to Evaluate Teachers and Schools*, (Ed. by J. Millman) Newbury Park, CA, Sage Publications. in press.

Weerasinghe, D., Orsak, T., and Mendro, R. (1997). *Value Added Productivity Indicators: A Statistical Comparison of the Pre-Test/Post-Test Model and Gain Model*. A paper presented at the 1997 Annual Meeting of the Southwest Educational Research Association, January, 1997, Austin TX.