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 statistical 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 lack of student 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 and Horn (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 5,197 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 Microsoft Excel's Ordinary Least Squares (OLS) software program to produce regression coefficients for each school. The deleted post-test scores were then estimated by HLM (see Bryk & Raudenbush, 1993), 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.

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

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 5,197 student data set was randomly reduced such that exactly 30 students were included per school. This created a new, reduced data file which contained 2,610 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 5,197 students were used in a fourth analysis. The initial exploratory analysis involved simple data analysis for the reduced data set.

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:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$
$$\beta_{1k} = \gamma_{10} + u_{1k}$$

Model 1B (OLS):

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

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

Model 2A (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\begin{aligned} \beta_{0k} &= \gamma_{00} + u_{0k} \\ \beta_{1k} &= \gamma_{10} + u_{1k} \\ \beta_{2k} &= \gamma_{20} + u_{2k} \\ \beta_{3k} &= \gamma_{30} + u_{3k} \end{aligned}$$

Model 2B (OLS):

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{MATH95}_{ik} + r_{ik}$$

The basic models with five student level variables and ten school level variables:

Model 3A (HLM):

Level 1:

$$\begin{aligned} \text{MATH96}_{ik} &= \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} \\ &+ \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{HISPANIC}_{ik} \\ &+ \beta_{4k}\,\text{BLACK}_{ik} + \beta_{5k}\,\text{GENDER}_{ik} \\ &+ \beta_{6k}\,\text{MATH95}_{ik} + r_{ik} \end{aligned}$$

Level 2:

$$\beta_{pk} = \gamma_{p0} + \sum_{k=1}^{10} \gamma_{pk} W_{kj} + u_{pk}$$

$p = 0, 1, 2, ..., 6.$

where

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

$\gamma_{00}, \cdots, \gamma_{011}$ = level-2 intercept/slopes to model all β_{0k}s,
$\gamma_{10}, \cdots, \gamma_{111}$ = level-2 intercept/slopes to model all β_{1k}s,
$\gamma_{20}, \cdots, \gamma_{211}$ = level-2 intercept/slopes to model all β_{2k}s,
u_{0k}, u_{1k}, u_{2k} = level-2 random effects for school k.

Model 3B (OLS):

$$\begin{aligned} \text{MATH96}_{ik} &= \beta_{0} + \beta_{1}\,\text{CEN-POV}_{ik} \\ &+ \beta_{2}\,\text{CEN-COL}_{ik} + \beta_{3}\,\text{HISPANIC}_{ik} \\ &+ \beta_{4}\,\text{BLACK}_{ik} + \beta_{5}\,\text{GENDER}_{ik} \\ &+ \beta_{6}\,\text{MATH95}_{ik} + r_{ik} \end{aligned}$$

For this study, the SEIs were calculated only from HLM, two level models. The models used for the calculations were as follows:

Model 1 (HLM):

Level 1:

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

Level 2:

$$\begin{aligned} \beta_{0k} &= \gamma_{00} + u_{0k} \\ \beta_{1k} &= \gamma_{10} + u_{1k} \end{aligned}$$

Model 2 (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\begin{aligned} \beta_{0k} &= \gamma_{00} + u_{0k} \\ \beta_{1k} &= \gamma_{10} + u_{1k} \\ \beta_{2k} &= \gamma_{20} + u_{2k} \\ \beta_{3k} &= \gamma_{30} + u_{3k} \end{aligned}$$

Model 3 (HLM):

Level 1:

$$\begin{aligned} \text{MATH96}_{ik} &= \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} \\ &+ \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{HISPANIC}_{ik} \\ &+ \beta_{4k}\,\text{BLACK}_{ik} + \beta_{5k}\,\text{GENDER}_{ik} \\ &+ \beta_{6k}\,\text{MATH95}_{ik} + r_{ik} \end{aligned}$$

Level 2:

$$\beta_{pk} = \gamma_{p0} + \sum_{k=1}^{10} \gamma_{pk} W_{kj} + u_{pk}$$

$p = 0, 1, 2, ..., 6.$

The SEI is given by

$$\text{SEI*} = \gamma_{00}\,.$$

Results

The main objective of this study was to determine an acceptable methodology for estimating missing student post-test scores within a school effectiveness analysis. In pursuing the main objective, it was also possible to determine the variability of school ranking based on estimated data. Missing data were estimated by either using HLM estimated values for each school or by OLS estimation within each school for the first two models. Thus, predicted values were not across district but within school. OLS criteria forced district-wide calculations in Model 3B when schools were encountered that where composed of one ethnic group. Correlations were calculated among the actual scores, the two estimated scores, and the average post-test scores per school for each percentage of data estimated. Correlations were also computed among the SEIs for each percentage of data estimated.

Model 1A & 1B

The following tables display the correlations among the original data scores, HLM estimated scores, OLS estimated scores and the school average post-test score, each table reflecting a different percentage of the original data deleted. Also displayed are the correlations among the original SEIs and the SEIs calculated with each of the three estimated data.

Model 1A (HLM):

Level 1:

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

Level 2:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$
$$\beta_{1k} = \gamma_{10} + u_{1k}$$

Model 1B (OLS):

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

Model 1 (HLM): SEI CALCULATION

Level 1:

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

Level 2:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$
$$\beta_{1k} = \gamma_{10} + u_{1k}$$

Table 3. 1% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8132	1.0000	
OLS	0.8132	0.9949	1.0000
AVG	0.5224	0.6406	0.6463

Table 4. 1% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9994	1.0000	
OLS--SEI	0.9994	1.0000	1.0000
AVG-SEI	0.9986	0.9988	0.9986

Table 5. 2% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7844	1.0000	
OLS	0.7802	0.9955	1.0000
AVG	0.4673	0.5551	0.5518

Table 6. 2% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9981	1.0000	
OLS--SEI	0.9981	0.9999	1.0000
AVG-SEI	0.9962	0.9974	0.9986

Table 7. 5% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8158	1.0000	
OLS	0.8167	0.9941	1.0000
AVG	0.3710	0.4713	0.4623

Table 8. 5% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9952	1.0000	
OLS--SEI	0.9953	0.9997	1.0000
AVG-SEI	0.9862	0.9927	0.9908

Table 9. 10% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8343	1.0000	
OLS	0.8350	0.9917	1.0000
AVG	0.3893	0.5101	0.4855

Table 10. 10% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9911	1.0000	
OLS--SEI	0.9915	0.9987	1.0000
AVG-SEI	0.9730	0.9875	0.9808

Table 11. 20% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7934	1.0000	
OLS	0.7956	0.9842	1.0000
AVG	0.3452	0.5152	0.4241

Table 12. 20% SEI Correlations

	ACT--SEI	*HLM-SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9794	1.0000	
OLS--SEI	0.9812	0.9928	1.0000
AVG-SEI	0.9405	0.9755	0.9480

The first HLM model examined, Model 1A, used MATH95 to predict MATH96 at the first level with no school-level conditioning variables. Tables 3, 5, 7, 9, and 11 show the correlations among the actual, HLM estimated, OLS estimated and average post-test scores which were 1%, 2%, 5%, 10% and 20% deleted. Note that as the percentage of data deleted increased, the correlation between the actual scores and HLM estimated scores ranged from 0.7844 to 0.8343 whereas the correlation between the actual scores and OLS estimated scores ranged from 0.7802 to 0.8350. The weakest correlations existed between the actual scores and the average school post-test values with a range of 0.3452 to 0.5224. No noticeable pattern existed between the HLM and OLS estimated score correlations to the percentage of data estimated. It was obvious that the HLM and OLS models produced nearly identical results as their estimated values were correlated at a minimal value of 0.9917. Also note that as the percentage of data estimated increased, HLM estimated values were more highly correlated to the average post-test score than the OLS estimated scores, an indication of HLMs shrinkage to the overall mean.

Tables 4, 6, 8, 10, and 12 indicate correlations of SEIs using data from the three estimation sources. As the percentage of estimated data increased, all correlations decreased. In this basic model, it was interesting to note OLS estimated data results had slightly higher correlations with the original SEIs in comparison to HLM estimated data, with the greatest difference at the 20% level (0.9812 versus 0.9794). Note that even the average school value produced correlations within the range of 0.9405 to 0.9986 depending on percentage of missing data.

Now the question of “which is best” in terms of prediction must be decided. Clearly, HLM produced estimates more closely related to the original data than OLS, but not so clear was why the SEIs of OLS were more closely related to the original data than HLM. Light will hopefully be shed on this situation as models become more complex.

Model 2

This next analysis introduced CEN-COL and CEN-POV into the previous model for the prediction of MATH96. CEN-COL represents the percentage of households within the student’s block who attended college. CEN-POV represents the percentage of households who fall below the poverty level.

Model 2A (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\begin{aligned}
\beta_{0k} &= \gamma_{00} + u_{0k}\\
\beta_{1k} &= \gamma_{10} + u_{1k}\\
\beta_{2k} &= \gamma_{20} + u_{2k}\\
\beta_{3k} &= \gamma_{30} + u_{3k}
\end{aligned}$$

Model 2B (OLS):

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{MATH95}_{ik} + r_{ik}$$

Model 2 (HLM): SEI CALCULATIONS

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$
$$\beta_{1k} = \gamma_{10} + u_{1k}$$
$$\beta_{2k} = \gamma_{20} + u_{2k}$$
$$\beta_{3k} = \gamma_{30} + u_{3k}$$

Table 13. 1% Predicted Data Correlations.

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8195	1.0000	
OLS	0.8079	0.9653	1.0000
AVG	0.5224	0.6392	0.5804

Table 14. 1% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9985	1.0000	
OLS--SEI	0.9992	0.9977	1.0000
AVG-SEI	0.9984	1.0000	0.9976

Table 15. 2% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7845	1.0000	
OLS	0.7771	0.9702	1.0000
AVG	0.4673	0.5536	0.5259

Table 16. 2% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9980	1.0000	
OLS--SEI	0.9977	0.9995	1.0000
AVG-SEI	0.9958	0.9967	0.9957

Table 17. 5% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8076	1.0000	
OLS	0.8058	0.9669	1.0000
AVG	0.3710	0.4763	0.4537

Table 18. 5% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9948	1.0000	
OLS--SEI	0.9946	0.9988	1.0000
AVG-SEI	0.9843	0.9915	0.9887

Table 19. 10% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8267	1.0000	
OLS	0.8140	0.9274	1.0000
AVG	0.3893	0.5186	0.4474

Table 20. 10% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9894	1.0000	
OLS--SEI	0.9854	0.9902	1.0000
AVG-SEI	0.9708	0.9866	0.9692

Table 21. 20% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7872	1.0000	
OLS	0.7540	0.9172	1.0000
AVG	0.3452	0.5169	0.3969

Table 22. 20% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9748	1.0000	
OLS--SEI	0.9185	0.9447	1.0000
AVG-SEI	0.9343	0.9703	0.8906

Tables 13, 15, 17, 19, and 21 include correlations between the actual, HLM estimated, OLS estimated and average post-test scores for the indicated percentage of data estimated. As the percentage of estimated data increased, the correlations range from 0.7845 to 0.8267 for HLM estimates and 0.7540 to 0.8140 for OLS estimates. In all percentages, HLM estimates were more correlated with the actual data than the OLS estimates, although the differences were extremely slight in one case (0.0018 difference). Again the weakest correlations were between the actual score and the average school post-test value with a range of 0.3452 to 0.5224. It can be noted that as the percentage of estimated data increases, the difference in correlations between HLM and OLS also increased.

Tables 14, 16, 18, 20, and 22 reflect the correlations of SEIs. Once more, as the percentage of estimated data increased, the correlations of SEIs decreased. HLM generated SEIs more correlated with the original SEIs than did OLS, which is in contrast to the first model. The greatest divergence occurred at the 20% level with a difference of 0.0563 while all others were of smaller deviations. Over more, the SEIs from average post-test scores correlated much lower than the estimates within a range of 0.9984 to 0.9343.

The "which is best" decision leans more clearly toward HLM in this particular model.

The third model analyzed included MATH95, CEN-COL, CEN-POV with the new variables of GEN, HIS, BLK, (where GEN represents student gender, HIS represents a Hispanic student and BLK represents a black student) to model MATH96. Ten school conditioning variables were also included in the HLM analysis at the school level. At this point difficulties were encountered in the OLS program in that numerous schools had populations of strictly one ethnic composition; thus it failed to generate estimates. HLM circumvented this predicament by generating estimates for all schools. OLS estimates were now generated across all schools, thus eliminating the problems encountered within schools.

Model 3

Model 3A denotes a true, two-level, hierarchical model with conditioning variables at the second level. This model was compared to the OLS Model 3B where OLS did not adjust for conditioning variables.

Model 3A (HLM):

Level 1:

$$\begin{aligned}\text{MATH96}_{ik} = \beta_{0k} &+ \beta_{1k}\ \text{CEN-POV}_{ik} \\ &+ \beta_{2k}\ \text{CEN-COL}_{ik} + \beta_{3k}\ \text{HISPANIC}_{ik} \\ &+ \beta_{4k}\ \text{BLACK}_{ik} + \beta_{5k}\ \text{GENDER}_{ik} \\ &+ \beta_{6k}\ \text{MATH95}_{ik} + r_{ik}\end{aligned}$$

Level 2:

$$\beta_{pk} = \gamma_{p0} + \sum_{k=1}^{10} \gamma_{pk} W_{kj} + u_{pk}$$

p = 0, 1, 2, . . . , 6.

Model 3B (OLS):

$$\begin{aligned}\text{MATH96}_{ik} = \beta_{0} &+ \beta_{1}\ \text{CEN-POV}_{ik} \\ &+ \beta_{2}\ \text{CEN-COL}_{ik} + \beta_{3}\ \text{HISPANIC}_{ik} \\ &+ \beta_{4}\ \text{BLACK}_{ik} + \beta_{5}\ \text{GENDER}_{ik} \\ &+ \beta_{6}\ \text{MATH95}_{ik} + r_{ik}\end{aligned}$$

Model 3 (HLM): SEI CALCULATION

Level 1:

$$\begin{aligned}\text{MATH96}_{ik} = \beta_{0k} &+ \beta_{1k}\ \text{CEN-POV}_{ik} \\ &+ \beta_{2k}\ \text{CEN-COL}_{ik} + \beta_{3k}\ \text{HISPANIC}_{ik} \\ &+ \beta_{4k}\ \text{BLACK}_{ik} + \beta_{5k}\ \text{GENDER}_{ik} \\ &+ \beta_{6k}\ \text{MATH95}_{ik} + r_{ik}\end{aligned}$$

Level 2:

$$\beta_{pk} = \gamma_{p0} + \sum_{k=1}^{10} \gamma_{pk} W_{kj} + u_{pk}$$

p = 0, 1, 2, . . . , 6.

Table 23. 1% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7731	1.0000	
OLS	0.7573	0.9683	1.0000
AVG	0.5224	0.5959	0.4746

Table 24. 1% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9903	1.0000	
OLS--SEI	0.9915	0.9470	1.0000
AVG-SEI	0.9842	0.9873	0.9883

Table 25. 2% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7466	1.0000	
OLS	0.7108	0.9613	1.0000
AVG	0.4673	0.5352	0.3865

Table 26. 2% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9832	1.0000	
OLS--SEI	0.9818	0.9802	1.0000
AVG-SEI	0.9720	0.9709	0.9701

Table 27. 5% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7811	1.0000	
OLS	0.7619	0.9548	1.0000
AVG	0.3710	0.4595	0.3068

Table 28. 5% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9818	1.0000	
OLS--SEI	0.9767	0.9812	1.0000
AVG-SEI	0.9731	0.9915	0.9887

Table 29. 10% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8182	1.0000	
OLS	0.8075	0.9455	1.0000
AVG	0.3893	0.5064	0.3818

Table 30. 10% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9776	1.0000	
OLS--SEI	0.9710	0.9718	1.0000
AVG-SEI	0.9620	0.9648	0.9592

Table 31. 20% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7779	1.0000	
OLS	0.7684	0.9316	1.0000
AVG	0.3452	0.5018	0.3190

Table 32. 20% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9503	1.0000	
OLS--SEI	0.9114	0.9447	1.0000
AVG-SEI	0.9175	0.9532	0.9154

Tables 23, 25, 27, 29, and 31 show correlations between the actual, HLM estimated, OLS estimated and average post-test scores for the indicated percentage of data estimated. As the percentage of estimated data increased, the correlations range from 0.7466 to 0.8182 for HLM estimates and 0.7108 to 0.8075 for OLS estimates. In all percentages, HLM estimates were more correlated with the actual data than the OLS estimates. Note again that as the model increased in complexity with the inclusion of more student variables and the addition of school level variables, the correlations decreased in comparison to previous models for the identical level of data estimated.

Tables 24, 26, 28, 30, and 32 reflect the correlations of SEIs. For the most part, as the percentage of estimated data increased, the correlations of SEIs decreased. HLM generated SEIs more correlated with the original SEIs than did OLS. The greatest divergence occurred at the 20% level with a difference of 0.0389 while all others were of smaller deviations. Moreover, the SEIs from average post-test scores correlated much lower than the estimates within a range of 0.9842 to 0.9175. Tests of correlations indicate all were significant.

These three models indicate that HLM is more suitable for estimating missing data than OLS or the average school score. This advantage must be gained by HLM's adjustments for school trends in comparison to overall trends for student scores. Investigations into HLM's ability to predict continued with repeated deletion estimations on the original data set.

The next phase of the investigation focused on twenty-five repeated deletion trials for each percentage of estimated data. The original 5,197 students were used in the computation of SEIs using only HLM estimates of the missing data. The SEIs generated by the twenty-five trials were compared individually to the original SEI and then the average of the twenty-five trials was compared to the original SEI for the complete data set. The model for this comparison was:

Model 4 (HLM):

Level 1:

$$\begin{aligned}\text{MATH96}_{ik} &= \beta_{0k} + \beta_{1k}\,\text{LEP}_{ik} \\ &+ \beta_{2k}\,\text{HISPANIC}_{ik} + \beta_{3k}\,\text{BLACK}_{ik} \\ &+ \beta_{4k}\,\text{GENDER}_{ik} + \beta_{5k}\,\text{MATH95}_{ik} + r_{ik}\end{aligned}$$

Level 2:

$$\beta_{pk} = \gamma_{p0} + \sum_{k=1}^{10} \gamma_{pk} W_{kj} + u_{pk}$$

$$p = 0, 1, 2, ..., 5$$

Table 33. SEI Correlations with Actual SEI

	AVG(25) vs. ACTUAL	*MAX Corr.*	*MIN Corr.*
1 %	0.9998	0.9989	0.9978
2 %	0.9998	0.9985	0.9977
5 %	0.9996	0.9966	0.9936
10 %	0.9994	0.9937	0.9867
20 %	0.9983	0.9837	0.9735

Table 33 denotes the correlations between the original SEI for the complete data set and the average of the SEIs for twenty-five trials, the maximum correlation between the original SEI and the individual trials as well as the minimum correlation between the individual trials and the original SEIs. The obvious main observation was as the percentage of data increases, the correlation between the actual SEI and estimated data SEI also decreased. Although the correlations remain quite high, an analysis of the ranks of the SEIs revealed changes of up to ten places in rank.

Conclusions

Several observations appear relevant based on this study. First, and perhaps most important, HLM estimates and OLS estimates are both similar to the original data up to approximately the 10% level whereas HLM estimates are more accurate to the original for greater percentages. This highlights the advantage of implementing HLM in educational data analysis when a greater percentage of data is missing. Second, SEIs with HLM estimates of missing data and OLS estimates of missing data are highly correlated when up to 10% of data is estimated for a relatively simple model without school level conditioning variables. This allows a choice of which method to choose for estimating missing data. Differences emerge as estimation models became more complex. The contradicting observation to the previous point is that HLM was able to generate estimates when full rank was not achieved within schools. For example, when students were all of one ethnicity within a school, OLS estimations failed for within school estimation. The alternative was to carry out OLS estimations across schools but it sacrifices potentially useful within-school information.

Future analyses are planned to formulate a test statistic that determines when the deviations of estimated scores from the actual scores are significant, and the deviations of school ranks from actual ranks are significant, along with investigations into the rank changes about their respective quartiles.

References

Bryk, A. S., & Raudenbush, S. W. (1993). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications.

Mendro, R. L., Webster, W. J., Bembry, K. L., & Orsak, T. H. (1994, October). *Applications of Hierarchical Linear Models in Identifying Effective Schools.* Paper presented at the Annual Meeting of the Rocky Mountain Educational Research Association, Tempe, AZ.

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

Orsak, T. H., Mendro, R.L., Webster, W. J., & Weerasinghe, Dash, (1996, April). *Empirical Difficulties in Using Hierarchical Linear Models for School*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Sanders, W. L., & Horn, S. P. (1993). *The Tennessee Value-Added Assessment System (TVAAS): Mixed Model Methodology in Educational Assessment*. Knoxville, TN: University of Tennessee.

Webster, W. J., Mendro, R. L., Bembry, K. L., & Orsak, T. H. (1994, October). *Alternative Methodologies for Identifying Effective Schools*. Paper presented at the Annual Meeting of the Rocky Mountain Educational Research Association, Tempe, AZ.

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