Multiple Linear Regression Viewpoints

A Publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression: the General Linear Model (MLR:GLM/SIG)

MLRV

Volume 24 • Number 1 • Fall 1997

Editorial Board

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the AERA Special Interest Group on Multiple Linear Regression: the General Linear Model (MLR:GLM/SIG) through the University of Missouri at St. Louis. *MLRV* abstracts appear in CIJE, the ERIC system, and microform copies are available from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106. *MLRV* is listed in the *EBSCO Librarians Handbook* and with the *FAXON* and *READMORE* subscription agencies. MLR:GLM/SIG information and a membership application form can be obtained by writing, FAXing (314-516-5784), Voice Mailing (314-516-5785), or e-MAILing (sspaner@umslvma.umsl.edu) the Executive Secretary. 1997-98 SIG membership and subscription fees are: Individual - $10 for one year, $18 for two years; Library/Agency - $20 per year; and Student - $5 for one year. Fee payment should be made payable to the **Multiple Linear Regression SIG** and sent to Steven Spaner, MLR:GLM/SIG Executive Secretary, 408 Marillac Hall, Division of Educational Psychology, University of Missouri - St. Louis, 8001 Natural Bridge Road, St. Louis, MO 63121-4499.

Measuring School Effects With Hierarchical Linear Modeling: Data Handling and Modeling Issues

Eugene P. Adcock, Ph.D., Prince George's County Public Schools, Maryland
Gary W. Phillips, Ph.D. National Center for Education Statistics, U.S. Department of Education

Because public schools do not randomly assign students and teachers across schools (methodological utopia), multilevel evaluation models which account for student and school contextual and practice variables in their natural settings provide the most rigorous means for empirically showing what is actually happening in school classrooms. Still, no statistical methodology can make up for faulty design or bad data. This article presents some important practical issues regarding data handling for multilevel analysis methodology. Also presented are important modeling design issues that need to be considered when applying hierarchical linear models (HLM) to the measurement of schools and for determining which factors impact the value schools add to students' achievement.

The statistical method chosen for an analysis is usually a function of two things: the question being asked and the nature of the recorded data. In the case of measuring school effects, HLM is a multivariate regression-like analysis technique that was developed specifically for use in school effects research. HLM allows the examination of associations among multi-level, nested data such as students within schools by estimating simultaneous linear equations at the student level within schools and the school level between schools. HLM models explain student and school variation in achievement scores, using both student- and school-level variables as explanatory variables, while accounting for the variance at each level. In the Prince George's County Public School district, the HLM model has been used to rank schools on their contribution to student achievement beyond those associated with student poverty, student mobility and school poverty (i.e., *Value-added Index)*, and HLM was also used to evaluate which factors contribute to the value added by schools (Adcock, 1995; Adcock, 1997).

Despite the tremendous potential for HLM to show how schools are doing and what can be done to make them better, the types of evaluation-quality data necessary to support the different levels of analysis – student, teacher, classroom, school, district — are not supported by the data handling practices of most public school districts. The fact that HLM is a non-experimental design involving the analysis of relationships among variables at multiple levels in the educational system makes the integrity of the data support system critical. Analysis of multilevel data must begin with an understanding of relationships among the lowest level variables, how unbiased higher level variables are constructed from lower level variables, and the relationships among the lower level and higher level variables (Cooley, Lloyd, and Mao, 1981).

After the multilevel evaluation design has been determined (e.g., HLM), the availability of specified student, classroom, school and district level evaluation-quality data is a real-life issue to the practical application to school effectiveness studies. The first section of this paper will address the issue of school district data support for multilevel evaluation designs and the second section will address modeling issues important to the successful application of the HLM model.

Section One

Data handling and data analysis are not distinctly different. Due to the increasing popularity of causal analysis and structural equation models (e.g., LISREL, AMOS, HLM) in school effects studies, the problems inherent in the multilevel nature of educational data are becoming more widely recognized (Bentler and Chou, 1988). School district data management systems and school district evaluation offices need to get in sync with the research, evaluation and accountability needs fulfilled by multilevel analysis models.

The formulation of explicit multilevel models with hypotheses about effects occurring at each level and across levels places important structural features and

demands on data. Expressing relationships among variables within a given level, and specifying how variables at one level influence relations occurring at another require a data processing system purposefully designed to support such innovative analysis methods. Because multilevel model analysis requirements of school district data are statistical in nature, it is the responsibility of school district evaluation offices to develop a relational database that can provide hierarchically structured and linked data — students, classrooms, schools, and district — for analysis by these powerful and important multilevel evaluation methods. From the perspective of a school district staff member responsible for fulfilling the data requirements of two large scale HLM school effects studies, the following data handling issues are identified among those which are important to the application of HLM analysis, reporting the nature of the analysis to colleagues, and supporting continued multilevel analysis studies:

1. taking control of variable definitions and parameters in determining the unit of analysis;

2. variable selection and measurement standards for evaluation-quality data vs. colleagues' "wish list" for inclusion of "crucial variables" in the analysis model; and

3. harvesting raw data from school district legacy system sources.

Unit of Analysis

Who is a student? Who is a teacher? What constitutes participation behavior, class size and student instructional cost? What is a program, a treatment, a school? Because one can not analyze below the data level that you observe, record, store and manage, it is vital that the unit of analysis parameters for measured predictor variables are established by statistical staff with a definitive vision and understanding of analysis. Once a plausible causal model has been defined, the structural equations implied by that model determine the appropriateness of a particular data analysis scheme. If the causal models are multilevel (e.g., HLM), then analysis will occur at the different levels for a complete understanding of the teaching and learning phenomena under investigation. In particular, the potential contribution of multilevel analysis is a function of recorded data on each individual's singular characteristics, experiences, behaviors, and achievements. Furthermore, since HLM analysis procedures take both student and school information into account simultaneously, it is important that data representing the same variables between these levels are consistent, linked and stable.

Multilevel evaluation models which account for student and school contextual and practice variables in their natural settings (e.g., HLM) provide a viable means of empirically showing what is actually happening in school classrooms. Students who are highly mobile and schools with highly mobile populations, for example, represent contextual variables which can be represented at both the student level (Student Mobility) and the school level (School Mobility). Likewise, teachers who have service years in a particular school (School Vested) and total service years in the district (System Vested) provide teaching experience information which naturally vary across schools. Rigorous variable specifications must rely upon an understanding of the school system source data structure and multilevel analysis requirements. These specifications enable the appropriate unit of analysis construction for individual student and individual teacher variables which can, in turn, be aggregated to higher classroom, school and district levels yielding consistent and stable estimates at each level.

Table 1 lists operational examples of how the Prince George's school district evaluation office fulfilled the requirements for evaluation-quality variables included in a recent HLM value-added study of 120 elementary schools (Adcock, 1997). This research study had two foci: the effects of personal characteristics and individual educational experiences on student learning, and how these relations are in turn influenced by classroom organization and the specific behavior and characteristics of the teachers within the school. Correspondingly, the data have a two-level hierarchical structure. The Level-1 units are the persons, who are nested within the Level-2 units of schools.

Table 1

School Year 1994-95 (SY95) HLM Value-Added Assessment Study
Partial List of Individual (Level 1) and Elementary School (Level 2) Variables

Variable	*Definition*	*Parameters*
Student (Level 1)	For the value-added study, student is a SY95 *Maryland School Performance Assessment Program* (MSPAP) eligible examinee with at least one scale score in the content areas of reading, mathematics or science.	"Student" is a *Research, Evaluation and Assimilation Database* (READ) warehouse system data element defined as a child who has an assigned PGCPS enrollment date and location, student number, race code and gender code.
Teacher (Level 1)	For the value-added study, elementary school teacher is a "core teacher" who is responsible for delivering the PGCPS curriculum in the six MSPAP test content areas (i.e., mathematics, science, social studies, reading, writing, and language arts).	**Core teacher** is a READ data element representing a school-based certificated "classroom" teacher employed on the last day of the school year and who has the assigned responsibility to provide students instruction and assign course grades in one or more of the core academic subject areas of language (reading, English, etc.), mathematics, science or social studies.
Class Size (Level 2)	The total number of students enrolled on the last day of the school year divided by the number of core teachers employed on the last day of the school year for each elementary school.	"Core Teacher" is a READ-defined data element: See Level 1 definition for "teacher" listed above. "Class Size" is a constructed class student-teacher ratio similar to that used by R. F. Ferguson (1991).
Teacher College Training (Level 2)	The average academic training index of the core teachers in a school. Seven point scale: 1=Bachelors, 2=Bachelors+30 course credit hours (cch), 3=Masters/Equivalent, 4=Masters+15(cch), 5=Masters+30(cch), 6=Masters+60(cch), 7=Doctorate.	Computed from the sum of teacher college training index divided by the number of core teachers employed on the last day of the school year for each elementary school.
Teacher Cost Per Student (Level 2)	The average salary of the core teachers employed at end-of-year (EOY) multiplied by the number of classroom teachers assigned (= the budgeted[1] number or the actual number of core teachers observed, whichever greater) divided by the total number of students enrolled in school at EOY for each elementary school.	Permanent teachers who are replaced by long-term substitute teachers at EOY required the following correction for computing the school's teacher salary (numerator): The average salary of the observed permanent core teaching staff is multiplied by the number of budgeted core teachers in each school.

[1] Pupil Accounting and School Boundary "*Class Size Report: 1994-95.*"

Enrollment Mobility: School (Level 2)	The average total number of days that SY95 *Maryland School Performance Assessment Program* (MSPAP) examinees were NOT enrolled in the school in which they began taking the SY95 MSPAP test for the past 3 years (SY93-SY95) based upon their most recent occurrence of continuous enrollment in that school.	Only the last continuous enrollment period is considered. No school transfers after the start of MSPAP administration date are considered. Continuous school enrollment (i.e., 0 Mobility) for 3 years is 540 days (i.e., 180 * 3 years) for the MSPAP school.
Enrollment Mobility: System (Level 2)	The average total number of days that SY95 MSPAP examinees were NOT enrolled in the PGCPS system for the past 3 years (SY93-SY95) dating back from the start of MSPAP administration date.	Note: continuous system enrollment (i.e., 0 Mobility) for 3 years is 540 days (i.e., 180 * 3 years) for any combination of schools in the system.
Teacher Service Years at MSPAP School (Level 2)	The average total number of years that core teachers employed at SY95 MSPAP schools "belonged" to that school based upon their most recent occurrence of continuous employment in that school.	Only the last continuous "belonging" period is considered.
Teacher Service Years in PG System (Level 2)	The average total number of years that core teachers have been employed as certified teachers in the PGCPS system based upon their most recent occurrence of continuous employment in the system.	Only the last continuous "belonging" period is considered.
% of MSPAP Examinees African-American (Minority) (Level 2)	The proportion of the total SY95 MSPAP examinee population who are African-American for each elementary school.	School aggregate means of Minority = 1 and Other = 0 are actually proportion values of study students who are African-Americans.
% Poverty Among MSPAP Examinees (Level 2)	The proportion of the SY95 MSPAP examinee population who are receiving a free or reduced lunch.	School aggregate means of Poverty = 1 and Non-Poverty = 0 are actually proportion values of study students who are eligible for Free/Reduced meal program.
% of MSPAP Examinees TAG (Level 2)	The proportion of the total SY95 MSPAP examinee population who are identified as "talented and gifted" by the TAG Office.	
Teacher Days Absent in SY95 (Level 2)	The proportion of days the core teachers employed at end-of-year (EOY) were absent during SY95 for each elementary school.	Computed from sum of teacher days absent divided by sum of days "belonging" to school for all end-of-year (EOY) core teachers.
Teacher Salary (Level 2)	The average core teacher salary in a school.	Computed from the sum of the teacher salary, divided by the number of core teachers at EOY in a school. SY95 "A" Scale Tables used for salaries.
Achievement Test Scale Score in Reading, Mathematics and Science (Level 2)	The school's average unweighted third and fifth grade student performance for SY95 MSPAP reading, mathematics and science content areas.	A few elementary or "combination" schools did not have both third and fifth grade levels. Cases deleted from content area school aggregation if missing test scale score.

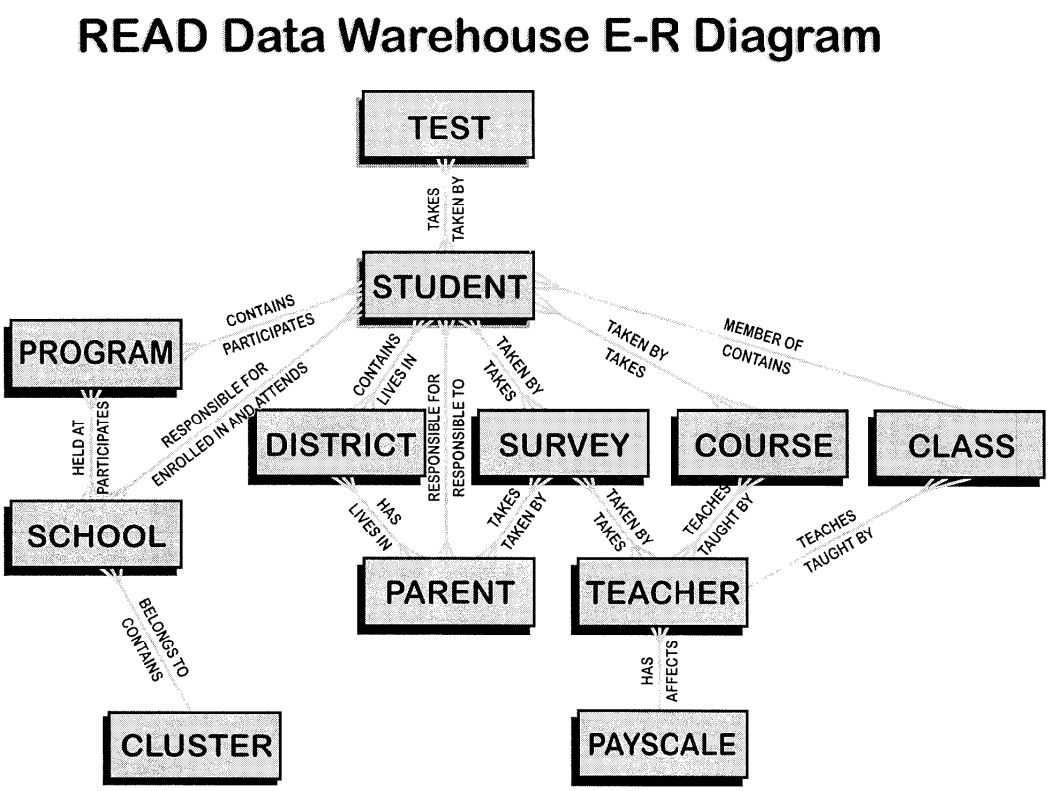


Figure 1

Data Entities And Their Relationships In The READ Warehouse

As can be seen from the list of variable in Table 1, selection of variables for this HLM research study was not limited to "available and easy" but included factors cited in school effects literature and by school policy members as important contributors to teaching and learning. Table 1 lists Level 1 variable definitions for student and teacher, and several Level 2 school aggregate variables used in a recent HLM study. The variable definitions and parameter specifications are also shown. Since Level 1 variables for individual characteristics, behaviors and achievements (e.g., Student SES, Student Mobility, and Teacher Training) are used to build Level 2 aggregate variable values, the Level 1 variables beyond "student" and "teacher" used in the study were omitted from the list because the reader can easily deduce the concomitant Level 1 definitions and specifications from those listed for Level 2.

Evaluation Variables vs. "Wish List"

You cannot analyze what you do not measure. It is around the conference table where evaluation study results are being presented that evaluation staff often learn from colleagues of the plethora of programs and initiatives which "explain everything!" but are missing from the causal evaluation model. For example, where are the: students' beginning achievement levels, gain scores, teacher inservice, Saturday Academies, parent participation, computer labs, dimension of learning instructional practices, content certified teachers, extra resource teachers,...,etc. in the multilevel model analysis? After all, schools are implementing one great thing after another great thing, and there is no measurement of these great things in the analysis model! Actually, there is no evaluation-quality measurement of these great practices at all, otherwise they would be in the analysis model. Statisticians have not been known to shy away from any available evaluation-

quality data that may correlate with student achievement.

As presented in the previous section, quality standards for evaluation analysis data must meet rigorous specifications. The evaluation-quality measurement standards include unit of analysis issues for case selection, assessment, scaling and recording. Often these evaluation-quality measurement standards are very difficult to achieve for many of the innovative practices, activities, experiences, and resources implemented by program staff and put forth as correlates of observed student achievement. In fact, when the measurement specifications are delineated for the inclusion of these practices (e.g., teacher inservice), program staff often find them too confining, burdensome, and in some cases menacing. Still, it is the responsibility of the school district evaluation office to provide guidance to staff interested in carrying out evaluation-quality measurement of their program's contribution to the value-added effects of schools.

Harvesting Data From Legacy Sources

" We do so much testing and surveys, plus filling out tons of data forms; how come we don't have any data for evaluating this program, that initiative or these schools?" With respect to research, the choice of data to analyze, debugging and preparation methods, management and storage procedures, and data layout is an act of theoretical preference (Davidson, 1996). A scientifically rigorous approach for research, evaluation and accountability has inherent data handling standards which sometimes render locally developed and administered data gathering information inadequate for evaluation purposes. Still, a database support system which can transform much of a school district's operational system data (e.g., course schedules, grades, tests, attendance, teacher service years, etc.) into a database system which meets the structural and statistical evaluation data standards for multilevel school and program evaluation studies is an indispensable tool for school district evaluation. In response to this vital need for pro-actively prepared evaluation-quality extant data on students, teachers, program/school participation measures, and resources the Research, Evaluation and Accountability staff of the PGCPS system has developed the *Research and Evaluation Assimilation Database (READ)* warehouse support system (Adcock, Haseltine, & Winkler, 1997). This comprehensive relational school district data warehouse model, READ, provides detailed achievement data together with contextual and process information at the various levels of students, classroom, teacher and schools. READ is well-suited for supporting scientifically rigorous multilevel HLM evaluation studies of student and school correlates with student achievement.

The READ data collection scheme focuses on collecting data for the following **five core database entities**: student, teacher, school, program and instructional finance. The READ warehouse sequential data processing procedures require data "scrubbing" for all incoming data. Scrubbing is a data warehouse term that includes the integration of legacy data from multiple sources and reformatting as necessary to ensure completeness, consistency, and accuracy. In addition, scrubbing data to evaluation requirement specifications often involves enhancement or derivation processing, partitioning and summarization of newly acquired legacy data. Transforming legacy data into evaluation-quality data is given such importance that the READ data warehousing pipeline has dedicated substantial resources to data verification, documentation, scrubbing and enhancement activities.

The design of the READ data warehouse follows logical relational database design with subject areas and their relationships. Figure 1 shows the Entity/Relationship Diagram (ERD) for the READ System's data warehouse.

In READ all input data is initially kept at the individual student (or teacher) level, and then aggregated at higher levels to meet complex evaluation data needs of multilevel analysis. Two-level HLM analysis, for example, may require the extraction of READ student level data for achievement, socio-economic status, ethnicity, etc., and school level data on teacher academic training, cost per student, mobility of student population, etc. The READ warehouse method of collecting, managing and extracting data permit this type of evaluation of the real-life multi-level nature of school district structure to be conducted. The next section describes some of the fundamental issues associated with modeling HLM analysis.

SECTION 2

This section is intended to provide researchers with the basic understanding of several statistical fundamentals of hierarchical linear models (HLM). We will introduce the HLM model, discuss centering, the estimation of school effects, and the empirical

Bayes estimation procedure. Along the way we will provide some practical advice in several other areas.

Simple versions of the HLM

To facilitate understanding we will illustrate all points with a simple 2-level HLM model with only one independent variable at both the student and school levels. We will also adopt the widely used notation provided by Bryk and Raudenbush (1992). At level I the dependent variable, Y_{ij}, will be math achievement and the independent variable, X_{ij} , will be socio-economic status (SES). At level II the independent variable will be the mean SES for school j, W_j.

Level I

$$Y_{ij} = ß_{0j} + ß_{1j} (X_{ij}) + r_{ij}, \qquad (1)$$

where

Y_{ij} = math achievement for student i in school j,

$ß_{0j}$ = expected math achievement. $ß_{0j}$ is an adjusted mean for school j such that

$$ß_{0j} = \mu_{y\,j} - ß_{1j} (X_{ij}),$$

$ß_{2j}$ = expected change in math achievement a unit change in X, and

r_{ij} = residual for student i in school j.

Level II

$$ß_{0j} = \gamma_{00} + \gamma_{01} (W_j) + \mu_{0j}, \qquad (2)$$

where

γ_{00} = predicted grand mean for math achievement for all schools based on W,

γ_{01} = change in expected school mean achievement ($ß_{0j}$) for a unit change in W,

μ_{oj} = unique effect of school j on the expected school achievement after controlling for W.

$$ß_{1j} = \gamma_{10} + \gamma_{11} (W_j) + \mu_{1j}, \qquad (3)$$

where

γ_{10} = SES slope for all schools.

γ_{11} = change in SES slope ($ß_{1j}$) for a unit change in W,

μ_{1j} = unique effect of school j on the SES slope after controlling for W.

At level I we make the assumptions that $E(r_{ij}) = 0$, and $Var(r_{ij}) = \sigma^2$. At level II we assume $E(\mu_{0j}) = E(\mu_{1j}) = 0$, $Var(\mu_{0j}) = \tau_{00}$, $Var(\mu_{1j}) = \tau_{11}$, $Cov(\mu_{0j},\mu_{1j}) = \tau_{01}$, and $Cov(\mu_{1j}, \mu_{0j}) = \tau_{10}$.

There are two important statistics that are based on these variances and covariances. The first is the intra-class correlation coefficient, p, (which indicates the overall degree of clustering within schools)

(1)

$$p = \tau_{00}/(\tau_{00} + \sigma^2), \qquad (4)$$

and the second is the reliability with which μ_{0j} is etimated by the ordinary least-squares estimate (OLS) $Y_{\cdot j} - ß_{1j} (X_{\cdot j})$ within each school

$$\lambda_j = \tau_{00}/(\tau_{00} + \sigma^2/n_j). \qquad (5)$$

The above first three equations can be expressed as a single level I model
by substituting equations 2 and 3 into 1. This yields the reduced form of the HLM model as follows

$$Y_{ij} = [\gamma_{00} + \gamma_{01} (W_j) + \mu_{0j}] + [\gamma_{10} + \gamma_{11} (W_j) + \mu_{1j}] (X_{ij}) + r_{ij}. \qquad (6)$$

As a general rule the coefficients of the level I model are treated as random while the level II (or the highest level in the model) are treated as fixed. Treating a level I coefficient as random indicates that the coefficient varies across schools (or level II units). One good way to better understand the HLM is to contrast it with other simpler models frequently used by education researchers. A number of commonly used simpler models can be obtained from equations 1-3 by fixing the level I parameters. For example, if there are no level I or level II independent variables then equation 1 becomes

$$Y_{ij} = ß_{0j} + r_{ij}, \quad (7)$$

and equation 2 becomes

$$ß_{0j} = \gamma_{00} + \mu_{0j}. \quad (8)$$

Substituting equation 8 into 7 yields

$$Y_{ij} = \gamma_{00} + \mu_{0j} + r_{ij}, \quad (9)$$

which is the one way analysis of variance (ANOVA) model. Another often used model can be derived from equations 1-3 by assuming no level II independent variable, and assuming that the $ß_{0j}$ at level I are fixed. When this is the case then equation 1 becomes

$$Y_{ij} = ß_{0j} + ß_{1j} (X_{ij} - X..) + r_{ij}, \quad (10)$$

equation 2 becomes

$$ß_{0j} = \gamma_{00} + \mu_{0j}, \quad (11)$$

and equation 3 becomes

$$ß_{1j} = \gamma_{10}, \quad (12)$$

which is the pooled within-school regression coefficient. Substituting equation 11 and 12 into 10 yields

$$Y_{ij} = [\gamma_{00}] + [\gamma_{10}](X_{ij} - X..) + \mu_{0j} + r_{ij}. \quad (13)$$

which is the analysis of covariance (ANCOVA) model (except for the fact that μ_{0j} is random instead of fixed).

Centering

Notice that in equation 8, X_{ij} was centered around the grand mean. In fact it is important to spend some time to be sure that the centering (especially at level I) is done in such a way that the interpretations of $ß_{0j}$ and γ_{00} are meaningful. There are essentially three ways to center at level I (uncentering, grand mean centering, and group mean centering) and two ways to center at level II (uncentering and grand mean centering). At level II, group mean centering and grand mean centering are really the same thing. Centering at level I determines the meaning of the Level I intercept and centering at level II determines the meaning of the level II intercept. In all cases the interpretation of the intercept is that it is the value of the dependent variable when the independent variable equals zero. In the following section we will only discuss centering at level I since the same interpretations apply to level II.

Uncentering

When X_{ij} is uncentered it means we wish to use the zero point in the original metric of X_{ij} as the defining point for $ß_{0j}$. In many areas of science the natural zero of X_{ij} has a practical interpretation. For example, if X_{ij} is the Celsius scale and Y_{ij} is the barometric pressure, then $ß_{0j}$ equals the barometric pressure when water freezes. In most situations in the social sciences there is not a natural zero point for X_{ij}. One notable exception to this is when dummy variable coding is used. For example, if $X_{ij} = 1$ for minority students and $X_{ij} = 0$ for non-minority students, then, $ß_{0j}$ equals the mean of Y_{ij} for non-minority students. If another dummy variable, Z_{ij}, is added to the level I equation, such as gender (where $Z_{ij} = 1$ for females and $Z_{ij} = 0$ for males), then $ß_{0j}$ equals the mean of Y_{ij} for non-minority males.

Group Mean Centering

In the social sciences the group mean of X_{ij} is often used as the zero point for X_{ij}. In group mean centering, $ß_{0j}$ equals the student's math achievement when $(X_{ij} - X._j)$ equals zero(which is at the group mean of X_{ij}). For example if X_{ij} is the SES of students, and Y_{ij} is the student's math achievement, then, $ß_{0j}$ equals the student's math achievement at the mean of SES. Another characteristic of group mean centering is that $ß_{0j}$ is always equal to the mean of Y_{ij}, or μ_{Yj}. Therefore, group mean centering is often used as the method of choice when the researcher is primarily interested in studying the variation in school means.

Grand Mean Centering

In the social sciences it is also common practice to center around the grand mean. An example of this was used in the above ANCOVA equation 13. In grand mean centering, $ß_{0j}$ equals the student's math

achievement when (X_{ij} - X..) equals zero (which is at the grand mean of X_{ij}). $ß_{0j}$ has a different interpretation in grand mean centering than it does in centering within groups. In grand mean centering $ß_{0j}$ is an adjusted mean such that $ß_{0j} = \mu_{Yj} + ß_{1j}$ (X_{ij} - X..). Grand mean centering is often used when the researcher is interested in estimating school effects.

Estimating School Effects

One of the main uses of HLM is to provide an index of school effectiveness. Once the school effects have been estimated then the researcher can rank schools on their effectiveness or use the effectiveness index as a dependent variable to investigate school factors that are related to effectiveness. A good example of school effects can be derived from the simple HLM model provided in equations 1-3. We rewrite these equations under the assumption that we use grand mean centering and for level I the intercept is random but the slope is fixed (i.e., constant across schools). Under these assumptions, equations 1-3 become

$$Y_{ij} = ß_{0j} + ß_{1j} (X_{ij}) + r_{ij}, \quad (14)$$

$$ß_{0j} = \gamma_{00} + \gamma_{01} (W_j) + \mu_{0j}, \quad (15)$$

$$ß_{1j} = \gamma_{10}. \quad (16)$$

Equations 14-16 are similar to the ANCOVA model except we have added W_j at level II. Substituting equations 15 and 16 into 14 yields the following reduced form

$$Y_{ij} = [\gamma_{00} + \gamma_{01} (W_j) + \mu_{0j}] + [\gamma_{10}](X_{ij} - X..) + r_{ij}.$$

Rearranging terms provides

$$\mu_{0j} = Y_{ij} - [\gamma_{00} + \gamma_{01} (W_j) + \gamma_{10} (X_{ij} - X..) + r_{ij}].$$

Averaging over student i within school j gives the estimate of school effects

$$\mu_{0j} = Y_{\cdot j} - [\gamma_{00} + \gamma_{01} (W_j) + \gamma_{10} (X_{\cdot j} - X..)]. \quad (17)$$

Empirical Bayes Estimation

In HLM the level I coefficients are usually estimated with an empirical Bayes procedure (Lindley and Smith, 1972). This procedure is different from the OLS used in most multiple regression procedures in that the level I estimates are weighted by the collateral estimates in level II. An example of this is found by inspecting more closely equations 1 and 2. We see that $ß_{0j}$ in equations 1 and 2 has two different OLS estimates, $ß^*_{0j}$

$$ß^*_{0j} = Y_{\cdot j} - ß^*_{1j} (X_{\cdot j}), \text{ and}$$

$$ß^*_{0j} = \gamma^*_{00} + \gamma^*_{01} (W_j).$$

The empirical Bayes estimate combines these two OLS estimates by weighting them according to the reliability,λ_j, of [$Y_{\cdot j}$ - $ß^*_{1j}$ ($X_{\cdot j}$)] as an estimate of $ß_{0j}$. The empirical Bayes estimate, $ß^{**}_{0j}$, is found by

$$ß^{**}_{0j} = \lambda_j[Y_{\cdot j} - ß^*_{1j} (X_{\cdot j})] + (1-\lambda_j)[\gamma^*_{00} + \gamma^*_{01} (W_j)]. \quad (18)$$

This approach was first introduced within the context of psychometrics by Kelley (1927). The weight λ_j is found by equation 5, and understanding this weight is key to appreciating the usefulness of the empirical Bayes estimation in HLM. As the reliability of the OLS estimate at level I approches unity, the best estimate of the within-school is from the data collected from within the school. However, as the reliability approaches zero (as when the number of students within the school is very low), then the best estimate of the within-school regression parameter is based on the regression parameters of similar schools within the system. The logic of this approach is identical to imputation in a survey sampling context. When data elements are missing for a school, a common practice is to substitute (or impute) data elements from similar schools to replace the missing value. Even treating the data as missing is the same as assuming that the missing data element is equal to the mean of the population.

The empirical Bayes estimate is an optimal estimate of $ß_{0j}$ in the sense that it has the smallest mean-squared error even though it is biased toward $\gamma^*_{00} + \gamma^*_{01}(W_j)$. The amount of bias is inversely related to λ_j. As a general rule the bias is negligible in schools with large sample sizes.

The empirical Bayes residual, μ^{**}_{0j}, is usually used by HLM researchers as the estimate of the school effect. Like the empirical Bayes estimate, μ^{**}_{0j}, is particulary biased in small schools. The relationship between the empirical Bayes residual and the OLS residual is as follows

$$\mu^{**}_{0j} = \lambda_j \mu^*_{0j}. \qquad (19)$$

As λ_j approaches zero, μ^{**}_{0j} also approaches zero. Even though the empirical Bayes residual is biased it is still considered by most educational researchers to be a better estimate than the OLS residual. This is because when the sample sizes are small the OLS residual will be unstable resulting in more chance occurences of extreme valus of μ^*_{0j}. Selecting out such extreme values of μ^*_{0j} for praise or blame will result in more false-positives and false-negatives than the empirical Bayes residual.

Authors' Note:
The discussion in this paper represents the views of the authors and does not represent those of the U.S. Department of Education.

References

Adcock, E. P. (1995). *Value-Added Effective Schools Study for Elementary Schools: 1994 Maryland School Performance Assessment Program Results*, Research, Evaluation & Accountability, Prince George's County Public Schools, MD., Research Report: 36-9-95.

Adcock, E. P. (1996). *Value-Added Assessment Study: Prince George's County Elementary Schools' 1995 MSPAP Results in Reading, Mathematics and Science*, (2nd Draft), Research, Evaluation & Accountability, Prince George's County Public Schools, MD., Research Report: 56-12-96.

Adcock, E. P., Haseltine, R., Winkler, L. H. (1997). *A Systemic Processing Model For Accountability*, Paper presented at the Panasonic Foundation Annual Partnership Conference IX, Colorado Springs, CO., March 1-3, 1997, Research Report: 60-3-97.

Arnold, C. L., (1995) *Using HLM and NAEP Data to Explore School Correlates of 1990 Mathematics and Geometry Achievement in Grades 4, 8, and 12: Methodology and Results*, National Center For Education Statistics (NAEP), Research and Development Report, Jan., 1995.

Bentler, P. M., & Chou, C. P. (1988). *Practical Issues in Structural Modeling*, Chapter 7 of *Common Problems/Proper Solutions: Avoiding Error in Quantitative Research*, Edited by J. Scott Long, Sage Publications, Newbury Park, CA.

Bryk, A.S., & Rondenbush, S.W., *Hierarchical Linear Models: Application and Data. Analysis Methods*, Sage, 1992.

Bryk, A. S., & Weisberg, H. I. Value-added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics*, 1976, 1, 127-155.

Cooley, W. W., Bond, L., Mao, B. J., (1981). *Analyzing Multilevel Data*, Chapter 3 of book entitled: *Educational Evaluation Methodology: The State Of The Art*, Edited by Ronald A. Berk, The Johns Hopkins University Press, Baltimore, MD.

Davidson, F. (1996). *Principles Of Statistical Data Handling*, Sage Publications, Inc., Thousand Oaks, CA.

Kelley, T.L. *The Interpretation of Educational Measurements*, New York: World Books, 1927.

Lindy, D.V. and Smith, A.F.M., *Bayes Estimates From the Linear Model, Journal of the Royal Statistical Society*, Series B, 1972, 34, pp. 1-41.

Phillips, G. W., & Adcock, E. P., (1997). Practical Applications of Hierarchical Linear Models to District Evaluations, *Journal of Multi-Linear Regression Viewpoints (*Spring, 1997).

Phillips, G. W., & Adcock, E. P., *Using Hierarchical Linear Models to Evaluate Schools*, Paper presented at American Educational Research Association, 1996 Annual Meeting, NY, Presentation 45.13, *Business Meeting/HLM Discussion/Symposium*, "Multiple Linear Regression: The General Model" (4/11/96).

Raudenbush, S. W., & Willms, J.D. The estimation of School Effects, *Journal of Educational and Behavioral Statistics*, Winter 1995, Vol. 20, No. 4, pp. 307-335.

Examples of Easily Explainable Suppressor Variables in Multiple Regression Research

Franklin T. Thompson and **Daniel U. Levine**
University of Nebraska at Omaha

Multiple regression techniques are a valuable tool in conducting ecological studies, especially when provisions are made to control for problems dealing with the interaction of variables. One problem in multiple regression research, the presence of suppressor variables, has the potential to seriously limit findings that can be reported, and in some cases may cause a researcher to pass over a useful data set. Researchers have long been aware of the presence of suppressors in multiple regression research, but there is little agreement as to why it exists or what to do about it. Several considerations in employing methods to "unsuppress" several data sets are discussed.

Suppressor variables have been defined as variables that substantially improve the prediction of a criterion through the addition of a variable which is uncorrelated or relatively little correlated with the criterion but is related to another predictor or set of predictors. When suppression occurs, addition of the suppressor to the regression equation frequently is associated with a sizable increase in the beta weight(s) of the previously suppressed predictor(s), and, in a forward stepwise analysis, an increase in R-square nearly as large or larger than that contributed by the previously-suppressed predictor. Given this pattern, one might well refer to the variable that thus "kicks up" the prediction as an "unsuppressor".

Although we have been examining and consuming research based on multiple regression for many years, we seldom have encountered studies incorporating or reporting clear (and valid) suppression effects. Analysis of the functioning of suppressor variables and their dynamics is still less frequent, even in research that could be clearly improved by devoting explicit attention to the effects and meaning of suppressor relationships.To illustrate the functioning of suppressors in actual studies, and ways in which analysis of their effects can enhance understanding of relationships in a data set, we will portray and summarize three examples of suppressor variables in multiple regression analysis. We will conclude with suggestions regarding procedures that can help researchers in determining how to proceed in multiple regression studies that examine or should include examination of suppressor relationships.

II. Education and Military Spending in 78 Nations

Our first example of suppression occurs in a data set that examines the relationships between spending for education and for the military (both assessed as percentage of gross national product) and average life expectancy in a diverse group of 78 nations. Using the 2 expenditure variables in a forward stepwise regression analysis to predict life expectancy, education enters first with a standardized coefficient of .3602 and an adjusted r square of .118. Military spending then enters with a standardized coefficient of -.364, the adjusted R square increases to .231, and the coefficient for education spending increases to .462. Thus education now has a stronger relationship with life expectancy than was true before controlling for military spending, and the explained variance has increased by .113 even though the zero-order correlation between military spending and life expectancy is only -.238.The addition of military spending to the analysis has unsuppressed the underlying pattern wherein education spending now is more strongly related to life expectancy than before, and the two predictors together explain more of the criterion variance than might have been expected from an examination of zero-order relationships.

Having noticed the appearance of suppressor dynamics, we examined what was taking place by calculating correlations between education spending and life expectancy in countries high and low in military spending, and by plotting this relationship while portraying the high/low level of military spending (Figure 1). The correlation analysis showed that among 42 nations with military spending below 3.5 percent of GNP, the correlation between education spending and life expectancy was .62; among 36 nations with military spending at or above 3.5 of GNP, the correlation was virtually non-existent at .02. Thus education spending is highly related to life expectancy in countries with relatively low spending devoted to military purposes, but not at all related to life expectancy in countries that have relatively high military expenditures.Given this pattern, it is intuitively easy to understand why taking account of military spending clarifies and enhances the effect of education spending in the regression analysis.

Examination of Figure 1 (which shows only a random .5 sub sample of the nations in the data set) further points to what may be happening. As shown in the plot, few countries that are high in military spending are very low in life expectancy, thus restricting possibilities for a high correlation between expectancy and other variables. Having identified these patterns, we can proceed to try to determine (not discussed in this paper) why nations that are high in military spending as a percent of GNP generally are not low in life expectancy, and how this situation may involve relationships between these and other variables.

III. Family Income and Academic Achievement in Two School Districts

Our second example involves analysis of relationships between a measure assessing family income (i.e., percent of students from low-income families) and average sixth-grade mathematics scores at 55 elementary schools in 2 school districts. The first variable to enter in predicting achievement in a forward stepwise regression analysis was the family income measure, which correlated at -.574 with achievement and accounted for an r square of .329 in the latter criterion. This correlation was not nearly as high as we generally have found in other analyses of achievement in large school districts.

The major reason for this relatively poor prediction became quickly apparent when a dummy variable portraying the 2 districts in the data set entered the multiple regression analysis, and when we plotted family income against achievement taking account of district (Figure 2). Although its zero-order correlation with achievement was only -.242, the dummy variable increased the R square to .625 and pushed up the regression coefficient for family income to -.874. As shown in Figure 2, family income is highly correlated with achievement in both districts but achievement in district 1 is generally higher than achievement in district 2.

Results were even more clear and dramatic when we combined total student achievement scores (combined math, reading, and language sub test scores) of 52 schools from the two districts and plotted them (Figure 3) against a poverty indicator we referred to as "school SES" (i.e., a factor analysis score made up of percent mobility, percent minority, and percent poor students). The zero-order relationship between achievement and school SES was .520, with an adjusted r square of .25. After once again controlling for district differences, the dummy variable increased the R square to .882 with 77% of the variance explained; a dramatic .52 increase in the adjusted r square at the .000 significance level (Table I). In addition to achievement being generally higher in district 1 than achievement in district 2, we are left to speculate that there may be additional influences (not discussed in this paper) differentiating the districts which help to further suppress the relationship between total achievement and our socioeconomic poverty variable.

When district-level achievement and other possible district differences are controlled through multiple regression analysis, the effects of family poverty and socioeconomic status are "unsuppressed", and we can proceed to additional analysis (not discussed in this paper) and research examining reasons for the high correlation with achievement, substantive possibilities for overcoming this association through improved instruction, and causes of differential achievement in the 2 districts.

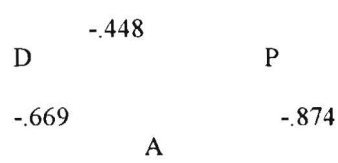
IV. Percentage of Students Residing Nearby and Math Achievement at 25 Schools in 1 School District

Various considerations led us to expect that schools which mostly enrolled students resident in their respective attendance areas in a school district we were studying would have proportionately lower achievement than schools which enroll higher proportions of students from distant neighborhoods. However, the correlation between percentage of resident students and average sixth-grade mathematics achievement was only -.023. Examination of the plot (Figure 4) suggested that a small group of 3 higher-than-predicted schools (i.e., box symbols with an x in Figure 4) was detracting from a clear relationship. We knew that reading scores accounted for more than 80 percent of the variance in math scores in this data set (as in many others), so we re-examined the relationship controlling for reading, and found that the standardized coefficient for percentage of resident students was now -.148. Inspection of the partial plot (Figure 5) indicated that increase in the size of the relationship between residency and math achievement was due to a reduction in the effects exercised by the three higher-than-predicted schools. Equally or more important, we were now in a better position to proceed with meaningful theoretical and quantitative exploration (not discussed in this paper) of relationships among variables in the analysis.

V. Discussion and Conclusions

The effects of a variable are "unsuppressed" when controlling for another variable indicates an increase in its relationship with the dependent variable. In the example involving family income and achievement described above and portrayed in Figure 2, the influence of poverty on achievement is increased to a multiple regression coefficient of -.874 from a zero-order correlation of -.574 because the latter relationship in a sense is a spuriously-low result of failure to control for district differences. As shown below, taking account of district in a path model helps the analyst understand underlying

relationships and computations. Let "D" stand for district, "P" for the poverty/family income measure, and "A" for achievement:



In this example, the zero-order correlation between P and A is the sum of the direct effect of P on A controlling for D and its indirect path through D. The calculations are as follows:

r pa = B pa.d + (r pd X B da.p)
-.574 = -.874 + (-.448 X -.669)
-.574 = -.874 + .300.
-.574 = -.574.

It is important to examine underlying interrelationships and even check out the calculations (as illustrated above) when one encounters regression data indicating that suppression effects are present. For one thing, the data produced by the computer may be invalid: If there is high multicollinearity among predictors or if there are too few cases to sustain valid computations given the number of predictors, multiple correlations and regression coefficients may invalidly indicate whopping increases as new relatively-poorly correlated variables are added to a stepwise multiple regression.

In addition, examining the model and/or the calculations can help the analyst understand the dynamics of forces at work in the data set. For example, examination of the model shown above underlines the fact that on the average, schools in District 1 (coded as 1) have higher poverty and achievement scores than schools in District 2 (coded as 2), even though the "normal" strong relationships between poverty and achievement are apparent within each district. Furthermore, these relationships are clearly visible in and, indeed, clearly suggested by the plot portrayed in Figure 2. These considerations help lead us to the following general conclusions:

1. Plotting relationships can be very helpful in understanding the dynamics of a data set including suppressors, and also in verifying that suppressor relationships actually are present. In some cases, plots can call attention to analytic possibilities not previously apparent that are worth further exploration.

2. Investigation of suppressor variables and relationships can greatly enhance analysis and understanding of what is occurring or may be implied in a researcher's data set. However, researchers should be cautious in identifying suppressors, because statistics pointing toward the presence of suppressors frequently are invalid indicators produced by a sample that is too small or by highly correlated predictors.

Table 1
Multiple Regression Analysis* of School Inputs Using Dependent Variable Achievement: Observing the Effects of a Suppressor Variable for Combined District Data

Step number	Independent variable	N	MR	Adj. R^2	Standard error	Beta	T score	p
1	School SES	52	.52	.25	.86	.52	- 4.27	.000
2	School SES					-.97	-12.29	.000
	District	52	.88	.77	.48	-.85	-10.74	.000

* Probabilities of F for entry = .05, and for removal = .10

The Use of the Johnson-Neyman Confidence Bands and Multiple Regression Models to Investigate Interaction Effects: Important Tools for Educational Researchers and Program Evaluators

John W. Fraas, Ashland University
Isadore Newman, The University of Akron

When investigating the impact of predictor variables on an outcome variable or measuring the effectiveness of an educational program, educational researchers and program evaluators cannot ignore the possible influences of interaction effects. The purpose of this paper is to present a procedure that educational researchers can follow in order to increase their understanding of the nature of the interaction effect between a dichotomous treatment variable and a continuous independent variable. This technique involves the use of three separate analytical techniques implemented in three steps. First, the interaction effect is statistically tested using a multiple regression model. Second, the interaction effect is plotted, and if the interaction effect is disordinal, the intersection point of the regression lines is calculated. Third, the Johnson-Neyman confidence limits are calculated. A list of the computer commands that can be used in conjunction with the SPSS/PC+ Statistics™ and the SPSS® for Windows™ computer software to calculate the Johnson-Neyman confidence limits is provided. In addition, this three-step analytical procedure is applied to a set of efficacy data that was collected in a study of the FOCUS instructional model in order to illustrate how it can be used by researchers and program evaluators.

Most educational researchers and program evaluators are aware of the need to investigate the possible existence of interaction effects. When an interaction effect is being examined, a researcher or an evaluator must answer two questions. First, what analytical technique can be used to test for the presence of an interaction effect? Second, what analytical technique can provide the maximum amount of information regarding the interaction effect when, in fact, it exists? Researchers and evaluators often consider the first question. The second question, however, appears to be a consideration less often. To obtain an in-depth understanding of the interaction effect, the researcher or evaluator must utilize an analytical technique that can provide such information. That is, the researcher must avoid a Type VI error (Newman, Deitchman, Burkholder, Sanders, & Ervin, 1976), which occurs when the analytical technique does not provide the appropriate or necessary information.

In this paper, we present a three-step analytical procedure for examining a linear interaction effect between a dichotomous treatment variable and a continuous independent variable. The first step in this analytical procedure, which was discussed in detail by McNeil, Newman, and Kelly (1996, pp. 127-140), requires the researcher to design models that are capable of statistically testing the interaction effect. The technique used in the second step, which was previously presented by Fraas and Newman (1977), Newman and Fraas (1979) and Pedhazur (1982, pp. 468-469), requires the researcher or program evaluator to calculate the point of intersection between the two regression lines. The third step requires that the Johnson-Neyman confidence bands be calculated. This technique has been discussed by Johnson and Neyman (1936), Rogosa (1980, 1981), Chou and Huberty (1992), and Chou and Wang (1992).

In this paper, we are stressing the importance of using these techniques together in a three-step analytical procedure. The use of this analytical procedure will provide researchers and program evaluators with the type of information that will increase their understanding of the nature of the interaction effect being examined. To illustrate the type of information that is produced by this three-step analytical procedure, we have analyzed the personal and teaching efficacy levels of teachers who were exposed to an instructional model developed by Russell (1992), which is referred to as FOCUS.

Analytical Technique Applied to Efficacy Scores

Even though Russell (1992) believed that the exposure to the FOCUS model would increase the participants' levels of personal and teaching efficacy, he was not willing to assume that those increases would be constant across the participants' pre-term efficacy levels. That is, when comparing the post-treatment personal efficacy and teaching efficacy scores of the teachers who were exposed to the FOCUS model to teachers who were not exposed to the model, the differences may not be consistent across the ranges of the pre-term efficacy scores. Thus, to understand the possible influence of the FOCUS model on the personal efficacy and teaching efficacy scores of teachers, it was essential, not only to test for the existence of pre-term efficacy scores by group interaction effects, but also to gain insight into the nature of these interaction effects, if in fact, they did exist.

Subjects

Sixty-eight teachers who were enrolled in graduate level classes offered by the Education Department of Ashland University were included in the evaluation of the FOCUS model. Ashland University is located in north-central Ohio, which contains rural, suburban, and urban school systems. The courses, which required 36 hours of instruction, were offered during a summer term. Twenty-nine of the 68 teachers were not exposed to the FOCUS model. These 29 teachers, who taught in grade levels that ranged from kindergarten to the twelfth grade, served as the Control Group. The other 39 teachers were exposed to the FOCUS model during the same academic summer term. These 39 teachers, who also taught in grade levels that ranged form kindergarten through the twelfth grade, were designated as the treatment group. This treatment group was referred to as the FOCUS Group.

Instruments

Various instruments are used to measure the level of a teacher's sense of efficacy. In this evaluation project, the Teacher Efficacy Scale, which was devised by Gibson and Dembo (1984), was used. This selection was consistent with the view expressed by Ross (1994) who stated in his extensive review of the teacher-efficacy research that:

> Future researchers should treat the [teacher efficacy] construct as a multi-dimensional entity rather than a singular trait, examining personal and general teaching efficacy separately rather than aggregating them [and they] should measure teacher efficacy with the most frequently used instruments to facilitate comparisons between studies (p. 27).

Each educator who participated in this study completed the Teacher Efficacy Scale at the beginning and end of the summer academic term. This instrument required each participant to rate each of 16 statements on a 1 (strong disagree) to 6 (strongly agree) scale. The ratings obtained from the first nine statements were summed to obtain a personal efficacy score for each teacher. A high score on these nine statements was interpreted to mean that the teacher had a high level of personal efficacy. And a low score would indicate that the teacher had a low level of personal efficacy. The other seven statements were used to measure a teacher's teaching efficacy score. The total score on these seven statements for each teacher was subtracted from 42. This procedure produced a teaching efficacy score that would be high for a teacher who had a high level of teaching efficacy. The score would be low for a teacher who had a low level of teaching efficacy.

Gibson and Dembo (1984) reported in their study that an analysis of internal consistency reliability values produced Cronbach's alpha coefficient values of .78 and .75 for the personal efficacy scores and teaching efficacy scores, respectively. In addition, Gibson and Dembo stated that a multitrait-multimethod analysis supported both convergent and discriminant validity of the instrument.

Hypotheses

Two null hypotheses were statistically tested in the efficacy study. These null hypotheses were as follows:

$1H_0$: The interaction effect between the pre-term personal efficacy scores and group membership does not account for some of the variation in the post-term personal efficacy scores.

$2H_0$: The interaction effect between the pre-term teaching efficacy scores and group membership does not account for some of the variation in the post-term teaching efficacy scores.

Each of these null hypotheses were statistically tested through the three step procedure presented in the following sections.

Step 1: Statistical Tests of the Interaction Effects

Step 1 of the three-step analytic procedure was implemented for the efficacy data by statistically testing multiple linear regression models that were designed to measure the linear interaction effects. As part of this hypothesis testing procedure, the data utilized in each model were tested for possible outlier values with tests of Cook's distance measures (Neter, Wasserman, & Kutner, 1985). Any person who had a value that would distort the regression analysis was reviewed to determine whether the data for that person should be eliminated. The test results of Cook's distance measures indicated that the data recorded for one teacher may distort the results obtained from the regression analysis of the teaching efficacy scores. After reviewing that teacher's data, the data were deleted from the regression analyses. Thus, a total of 68 teachers and 67 teachers were included in the regression analyses of the personal efficacy scores and teaching efficacy scores, repectively.

The model that was designed to test 1Ho, which dealt with the teachers' personal efficacy scores, contained three independent variables. The teachers' post-term personal efficacy scores served as the dependent variable for this model. One of the independent variables included in this model consisted of the teachers' pre-term personal efficacy scores. This variable was labeled Pre-Term PE. The second independent variable included in this model was the Group variable. This Group variable consisted of the values of zero and one. A value of one indicated that the teacher was in the FOCUS Group, and a zero value meant that the teacher was in the Control Group. The third variable included in this model was formed by multiplying the Pre-Term PE variable by the Group variable. The inclusion of this variable, which was labeled (Pre-Term PE)*(Group), allowed us to use the regression model to calculate the difference between the slopes of the Control and FOCUS groups' regression lines.

The *t*-test value of the regression coefficient for the (Pre-Term PE)*(Group) variable was used to test $1H_0$. Since this study involved two dependent variables, i.e., the personal efficacy and teaching efficacy variables, the alpha level for the t test of this regression coefficient value was set at .025, which is equal to .05 divided by 2. The chance of committing a type I error was reduced by using this alpha value (Newman & Fry, 1972).

The results obtained from the analysis of the regression model are contained in Table 1. The t test of regression coefficient for the (Pre-Term PE)*(Group) variable (t = -2.44, p = .0175) indicated that the difference between the slopes of the regression lines of the FOCUS and Control groups was statistically significant at the .025 level, that is, $1H_0$ was rejected. Thus, the differences between the post-term personal efficacy scores of the FOCUS and Control groups were not constant across the range of pre-term personal efficacy scores.

Table 1
Regression Results for the Post-Term Personal Efficacy Scores

Regression Model

Variable	Regression Coefficient	t Value	p Value
(Pre-Term PE)*(Group)	-0.538	-2.44	0.018
Pre-Term PE	0.852	5.17	<.000
Group	25.124	2.87	0.006
Constant	6.362	0.97	0.338
R2 = .370			
Adjusted R2 = .341			
N = 68			
Residual Sum of Squares = 2495.58			

Note. The values for the Group variable are zero and one for teachers in the Control and FOCUS groups, respectively.

Table 2
Regression Results for the Post-Term Teaching Efficacy Scores

Regression Model

Variable	Regression Coefficient	t Test Value	p Value
(Pre-Term TE)*(Group)	0.703	2.742	0.008
Pre-Term TE	0.153	0.79	0.433
Group	-14.569	-2.339	0.023
Constant	19.8	4.331	<.000
R2 = .347			
Adjusted R2 = .316			
N = 67			
Residual Sum of Squares = 1334.318			

Note. The values for the Group variable are zero and one for teachers in the Control and FOCUS groups, respectively.

The teaching efficacy scores served as the dependent variable in the regression model that was used to test 2Ho. Similar to the previous regression model, this model included three independent variables. One of these independent variables was composed of the teachers' pre-term teaching efficacy scores. This variable was labeled Pre-Term TE. A second independent variable included in the model was the Group variable. The third independent variable included in the model was generated by multiplying the Pre-Term TE variable by the Group variable. This variable, which was labeled (Pre-Term TE)*(Group), was used to estimate the difference between the slopes of the regression lines for the Control and FOCUS groups.

The values generated by the analysis of the regression model used to test $2H_0$ are listed in Table 2. The t test of the regression coefficient for the (Pre-Term TE)*(Group) variable (t = 2.742, p = .008) indicated that the interaction effect was statistically significant at the .025 level. Thus, the differences between the post-treatment teaching efficacy scores of the FOCUS and Control groups were not constant across the range of pre-term teaching efficacy scores.

Step 2: Calculation of the Point of Intersection

The second step of the three-step analytical procedure was implemented by, first, graphing each of the interaction effects. If a given the interaction effect is disordinal, the point of intersection between the two regression lines would be calculated. If the interaction effect is ordinal, that is, the regression lines do not intersect in the relevant range, the researcher would proceed to Step 3.

The interaction effect between the Pre-Term PE variable and the Group variable is diagramed in Figure 1. Since the interaction effect was disordinal, the point at which the two regression lines intersected was calculated as follows:

1. The value of zero was substituted for the Group variable in the regression equation contained in Table 1 to obtain the regression line for the Control Group.

Y = 6.362 - .538*(Pre-Term PE)*(Group) + .852*(Pre-Term PE) + 25.124*(Group)

Y = 6.362 - .538*(Pre-Term PE)*(0) + .852*(Pre-Term PE) + 25.124*(0)

Y = 6.362 + .852*(Pre-Term PE)

2. The value of one was substituted for the Group variable in the regression equation contained in Table 1 to obtain the regression line for the FOCUS Group.

Y = 6.362 - .538*(Pre-Term PE)*(Group) + .852*(Pre-Term PE) + 25.124*(Group)

Y = 6.362 - .538*(Pre-Term PE)*(1) + .852*(Pre-Term PE) + 25.124*(1)

Y = 31.486 + .314*(Pre-Term PE)

3. The two regression lines were set equal to each other and the researcher solved the equation for Pre-Term PE.

6.362 + .852*(Pre-Term PE) = 31.486 + .314*(Pre-Term PE)

.538*(Pre-Term PE) = 25.124

Pre-Term PE = 46.7

As indicated by the results of this calculation and the graph of the disordinal interaction effect contained in Figure 1, the post-term personal efficacy scores of the teachers in the FOCUS Group were higher than the post-term personal efficacy scores of the teachers in the Control Group when their pre-term personal efficacy scores were less than 47. The post-term personal efficacy scores of the teachers in the Control Group, however, were higher than the post-term personal efficacy scores of the teachers in the FOCUS Group when their pre-term personal efficacy scores were greater than or equal to 47.

The interaction effect between the Pre-Term TE variable and the Group variable, which is diagramed in Figure 2, was also disordinal. Using the values produced by the regression analysis contained in Table 2, the point at which the two regression lines for the post-term teaching efficacy scores intersected was calculated in the same manner as was the intersection point for the personal efficacy scores. The calculations were as follows:

1. The value of zero was substituted for the Group variable in the regression equation contained in Table 2 to obtain the regression line for the Control Group.

Y = 19.800 + .703*(Pre-Term TE)*(Group) + .153*(Pre-Term TE) - 14.569*(Group)

Y = 19.800 + .703*(Pre-Term TE)*(0) + .153*(Pre-Term TE) - 14.569*(0)

Y = 19.800 + .153*(Pre-Term TE)

2. The value of one was substituted for the Group variable in the regression equation contained in Table 2 to obtain the regression line for the FOCUS Group.

Y = 19.800 + .703*(Pre-Term TE)*(Group) + .153*(Pre-Term TE) - 14.569*(Group)

Y = 19.800 + .703*(Pre-Term TE)*(1) + .153*(Pre-Term TE) - 14.569*(1)

Y = 5.231 + .856*(Pre-Term TE)

3. The two regression lines were set equal to each other and the researcher solved the equation for Pre-Term TE.

19.800 + .153*(Pre-Term TE) = 5.231 + .856*(Pre-Term TE)

.703*(Pre-Term TE) = 14.569

Pre-Term TE = 20.7

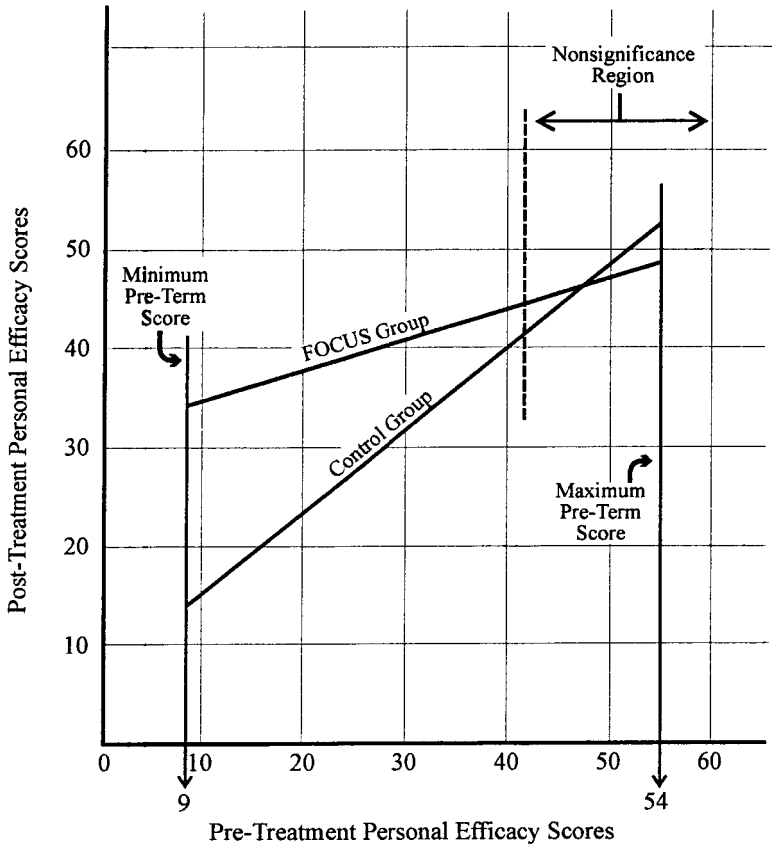


Figure 1. Pre-Term Personal Efficacy Scores by Group Interaction.

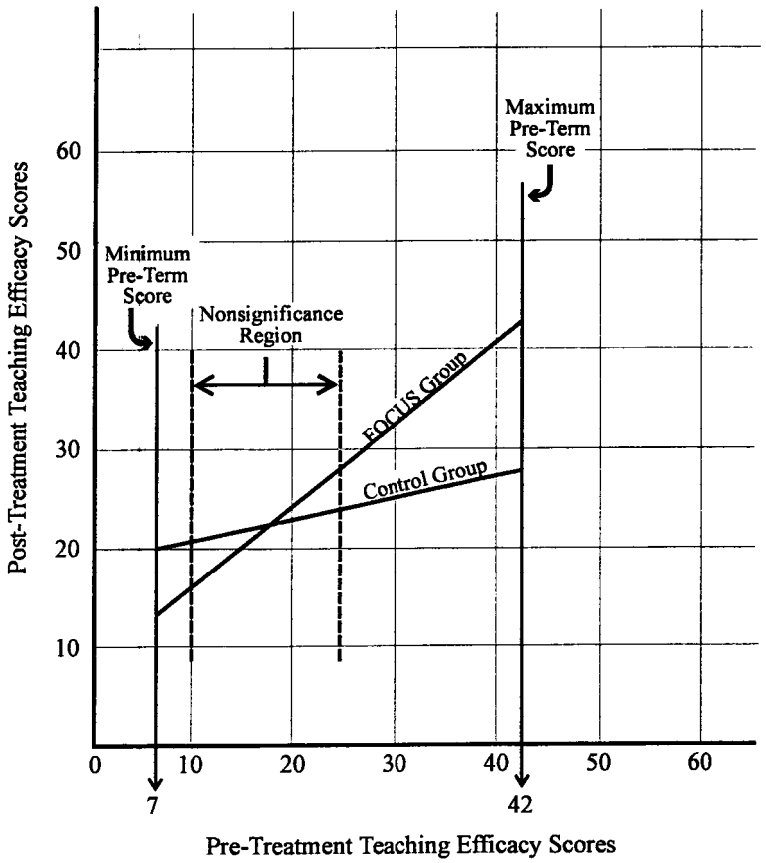


Figure 2. Preterm Teaching Efficacy Scores by Group Interaction

Table 3

Percentage of Teachers with Pre-Term Efficacy Scores Located In Various Regions Above and Below the Points of Intersection Between the Two Pairs of Regression Lines

Post-Term Personal Efficacy Scores	Post-Term Teaching Efficacy Scores	
	FOCUS > Control	FOCUS < Control
FOCUS > Control	72%	19%
FOCUS < Control	6%	3%

The post-term teaching efficacy scores of the teachers in the Control Group were greater than the post-term teaching efficacy scores of the teachers in the FOCUS Group when their pre-term teaching efficacy scores were below 21. In addition, the post-term teaching efficacy scores of the teachers in the FOCUS Group were greater than the post-term teaching efficacy scores of the teachers in the Control Group when their pre-term teaching efficacy scores were greater than or equal to 21.

After the intersection point is calculated in a study that investigates an interaction effect between a continuous independent variable and a treatment variable, it is important to note the percentage of the study's participants who have scores above and below the intersection point. For the efficacy data of the 67

teachers who were included in both analyses, the percentages are listed in Table 3. As indicated in Table 3, 72% of the teachers had pre-term efficacy scores that corresponded to points on the regression lines where the teachers had higher post-term personal efficacy scores and higher post-term teaching efficacy scores when exposed to the FOCUS model. Only 3% of the teachers had pre-term efficacy scores that corresponded to points on the regression lines where the teachers had lower post-term personal efficacy scores and lower post-term teaching efficacy scores when exposed to the FOCUS model. Nineteen percent of the teachers had pre-term efficacy scores that corresponded to points on the regression lines where the teachers had higher post-term personal efficacy scores and lower post-term teaching efficacy scores when exposed to the FOCUS model. And 6% of the teachers had pre-term efficacy scores that corresponded to points on the regression lines where the teachers had lower post-term personal efficacy scores and higher post-term teaching efficacy scores when exposed to the FOCUS model.

With respect to these percentages, It is important to realize that the differences between the post-term efficacy scores of the FOCUS and Control groups may be statistically significant only for certain ranges of the pre-term efficacy scores. Thus, before conclusions are drawn with respect to who benefits and who does not benefit from being exposed to the FOCUS model, it is essential to determine the ranges of pre-term efficacy scores in which the differences between the post-term efficacy of the teachers in the FOCUS Group and the teachers in the Control Group are statistically significant. Step 3 of this three-step analytical procedure is designed to determine these statistically significant ranges.

Step 3: Calculation of the Johnson-Neyman Confidence Bands

The third step of the three-step analytical procedure requires that the Johnson-Neyman confidence limits be calculated for each statistically significant interaction effect. It should be noted that some researchers have argued that the Johnson-Neyman regions of significance are non simultaneous ones (Potthoff, 1964 and Rogosa, 1980, 1981). Based on empirical results by Chou and Huberty (1992) and Chou and Wang (1992), it appears that the Johnson-Neyman technique can be used to make simultaneous inferences provided that the slope homogeneity assumption is statistically tested and rejected. Since $1H_0$ and $2H_0$ were rejected, it was appropriate to calculate Johnson-Neyman (1936) confidence bands for the nonsignificance regions for the efficacy scores.

The program that was used to calculate the Johnson-Neyman confidence bands, which can be used in conjunction with the SPSS/PC+ Statistics™ software (SPSS Inc., 1990) and the SPSS® Base 7.0 for Windows™ (SPSS Inc., 1996), is listed in the Appendix. The program, which calculates the Johnson-Neyman significance bands as suggested by Pedhazur (1982, pp. 169-171), requires that 12 values be provided. A description of the required values, as well as their labels, are as follows:

1. The symbol ss1 represents the pre-term sum of squares value for the Control Group.
2. The symbol ss2 represents the pre-term sum of squares value for the FOCUS Group.
3. The symbol n1 represents the sample size of the Control Group.
4. The symbol n2 represents the sample size of the FOCUS Group.
5. The symbol sumresid represents the residual sum of squares value of the regression model.
6. The symbol mean1 represents the mean of the pre-term scores of the Control Group.
7. The symbol mean2 represents the mean of the pre-term scores of the FOCUS Group.
8. The symbol slope1 represents the slope of the regression line for the Control Group.
9. The symbol slope2 represents the slope of the regression line for the FOCUS Group.
10. The symbol int1 represents the intercept point of the regression line for the Control Group.
11. The symbol int2 represents the intercept point of the regression line for the FOCUS Group.
12. The symbol *fcrit* represents the critical F value with 1 and N - 4 degrees of freedom.

The sum of squares values, the sample sizes, and the mean values can be obtained from the printout generated by the DESCRIPTIVE subprogram of the SPSS/PC+ STATISTICS™ software (SPSS Inc., 1990) or the SUMMARIZE subprogram of the SPSS® Base 7.0 for Windows™ software (SPSS Inc., 1996), with each of the two groups being analyzed separately. The residual sum of squares value, the slope values, and the intercept-point values can be obtained from the printouts generated by the REGRESSION subprogram of either the SPSS/PC+ STATISTICS™ software or the SPSS® Base 7.0 for Windows™ software. The critical F value can be obtained from an F-Distribution Table.

The data line of the program listed in the Appendix, which utilized the freefield format, contains the data used to generate the Johnson-Neyman confidence limits for the personal efficacy scores. The data line used for the analysis of the teaching efficacy scores was as follows: 567.30 745.82 29 38 1334.32 23.24 24.71 .15 .86 19.80 5.23 4.00. Note that the numerator degrees of freedom (df_n) and the denominator degrees of freedom (df_d) values were 1 and 64 (68-4), respectively, for the analysis of the post-term personal efficacy scores. For the analysis of the post-term teaching efficacy scores, the values for df_n and the df_d were 1 and 63 (67-4), respectively. In addition, the confidence level was set at .95 for each set of limits.

The upper limit for the 95% confidence bands for the personal efficacy scores was 81.8, which was above the maximum score of 54 points on the personal efficacy section of the Teacher Efficacy Scale. The lower limit was 40.7. Based on these limits, which are included in Figure 1, it can be concluded that the post-term personal efficacy scores for the teachers in the FOCUS and Control groups were not statistically significantly different when their scores were greater than or equal to 41. The post-term personal efficacy scores of the teachers in the Focus Group were statistically significantly higher than the corresponding scores of the teachers in the Control Group, however, when their pre-term scores were less than 41.

The lower limit of the 95% Johnson-Neyman confidence limits for the regression lines diagramed in Figure 2 was equal to 9.97, which was less than three points above the minimum score of 7 that a teacher could receive on the teaching efficacy section of the Teacher Efficacy Scale. It should be noted, however, that none of the teachers included in this analysis had a pre-term teaching efficacy score below 13. Thus, none of the teachers included in this study had a score below the lower limit of the nonsignificance region. The upper limit of the nonsignificance region of the Johnson-Neyman 95% confidence limits for the pre-term teaching efficacy scores was 23.8. Thus, the post-term teaching efficacy scores of the teachers in the FOCUS and Control groups were not statistically significantly different when their pre-term teaching efficacy scores were less than 24. The post-term teaching efficacy scores of the teachers in the FOCUS Group, however, were statistically significantly higher than the post-term teaching efficacy scores of the teachers in the Control Group when their pre-term teaching efficacy scores were equal to or greater than 24.

To understand the implications of the nonsignificant regions as well as the significant regions for the two sets of regression lines, it is important to note the location of the teachers' pre-term efficacy scores along the two sets of regression lines. As indicated by the percentages contained in Table 4, 31% of the teachers who were included in both regression analyses had pre-term efficacy scores that corresponded to points on the regression lines where the post-term efficacy scores of the teachers in the FOCUS Group were statistically significantly higher than the scores of the teachers in the Control Group on both efficacy scales. In addition, 42% of the teachers had pre-term efficacy scores that corresponded to points on the regression lines where the post-term efficacy scores of the teachers in the FOCUS Group were statistically significantly higher than the scores of the teachers in the Control Group on one of the two efficacy scales. The remaining 27% of the teachers had pre-term efficacy scores that corresponded to points on the regression lines where the post-term efficacy scores of the two groups were not statistically significantly different on either efficacy scale.

Table 4

Percentage of Teachers with Pre-Term Efficacy Scores Located in the Various Significant and Nonsignificant Regions

Post-Term Personal Efficacy Scores	Post-Term Teaching Efficacy Scores		
	FOCUS > Control	FOCUS = Control	FOCUS < Control
FOCUS > Control	31%	21%	0%
FOCUS = Control	21%	27%	0%
FOCUS < Control	0%	0%	0%

Thus, a total of 73% had pre-term efficacy scores that were located at points on the regression lines where the post-term efficacy scores of the teachers in the FOCUS Group were statistically significantly higher than the post-term efficacy scores of the teachers in the Control Group on at least one of the two efficacy scales. None of the teachers (0%) had pre-term efficacy scores that were located at points on the regression lines where the post-term efficacy scores of the teachers in the Control Group were statistically significantly higher than the post-term efficacy scores of the teachers in the FOCUS Group on either of the two efficacy scales.

Implications Based on the Results of the Three-Step Analytical Procedure.

It is important to understand what each step in this three-step analytical procedure reveals about the linear interaction effects. The results of Step1 indicate that both interaction effects were statistically significant. A more in-depth understanding of these interaction effects, however, is
obtained by reviewing the information generated by Steps 2 and 3 of this three-step analytical procedure.

The graphs containing the interaction effects and the points of intersection between the regression lines for the personal efficacy scores and the teaching efficacy scores, which were completed in Step 2, revealed that both interaction effects were disordinal and the regression lines for the personal efficacy scores and the teaching efficacy scores intersected at 46.7 and 20.7, respectively. These graphs and the intersection points appear to suggest that, with respect to their post-term efficacy scores, certain teachers would benefit from being exposed to the FOCUS model, while exposure to the FOCUS model would be detrimental to other teachers. In addition, these points of intersection could possibly be used to identify which teachers would and would not benefit from exposure to the FOCUS model. Before such a conclusion is reached, however, it is important to realize that the differences between the post-term efficacy scores of the teachers in the FOCUS and Control groups, who have pre-term scores near the intersection points, could simply be due to noise or random variation. That is, the post-term scores of the students in the two groups are statistically significantly different only for pre-term scores that are located some distance above and below the intersection points. Thus, before one should draw a conclusion with respect to the nature of these interaction effects, it is essential to review the information provided by the Johnson-Neyman confidence limits calculated in Step 3.

The significance region between the two regression lines that were designed to analyze the
post-term personal efficacy scores included only the pre-term personal efficacy scores that were less than 41. In addition, the significance region between the two regression lines that were designed to analyze the post-term teaching efficacy scores included only the pre-term teaching efficacy scores that were greater than or equal to 24. Thus, as indicated by the interaction effects contained in Figures 1 and 2, whenever the post-term efficacy scores of the two groups were statistically sigficantly differerent, the post-term efficacy scores of the Focus Group exceeded the post-term efficacy scores of the Control Group.

Thus, a majority of teachers (73%) had pre-term efficacy scores that placed them in ranges along the regression lines that indicated that the post-term efficacy scores of the teachers in the Focus Group, on at least one of the efficacy scales, were statistically significantly higher than the post-term efficacy scores of the teachers in the Control Group. It is important to also note that in spite of the fact that the interaction effects were disordinal, the reverse statement is not true. That is, none of the teachers had pre-term efficacy scores in the ranges along the regression lines that indicated that the post-term efficacy scores of the Focus Group were statistically significantly lower than the post-term efficacy scores of the Control Group on either of the two efficacy scales. The remaining 27% of the teachers had pre-term efficacy scores in the ranges
along the regression lines that indicated that the post-term efficacy scores of the FOCUS and Control groups were not statistically significantly different on either of the two efficacy scales.

Based on this information, one would not use the intersection points between the regression lines to determine who would and who would not benefit from being exposed to the FOCUS model. Rather, it would be more appropriate, keeping in mind research design limitations, to suggest that, based on pre-term efficacy levels, exposing the teachers to the FOCUS model would be beneficial to the majority of teachers and it would not be detrimental to any one group of teachers. Educational researchers and program evaluators would reach this conclusion only by using this three-step analytical procedure.

Summary

It is important for educational researchers and program evaluators to increase their understanding of the interaction effects that may be present in their data. We believe that a more in-depth understanding

of a linear interaction effect between a continuous independent variable and a dichotomous treatment variable can be obtained if the educational researcher or program evaluator follows the three-step analytical procedure that was presented in this paper.

Two points should be noted regarding this three-step analytical procedure. First, the use of a multiple regression model to statistically test the interaction effect, which is undertaken in Step 1, is an essential analytical procedure to consider when investigating the difference between the scores of two groups. This test of the homogeneity of the slopes of the regression lines allows the researcher to not only to determining if the interaction effect is statistically significant, but it also permits simultaneous inferences to be made from the Johnson-Neyman confidence bands, which are calculated in the third step of this analytical procedure.

Second, the calculation of the intersection point between the two regression lines in Step 2 could posssibly provide a researcher or program evaluator with information that could be used to identify groups of people who would benefit from being exposed to the treatment being investigated. It is important to realize, however, that the difference between the post-term scores of the students in the two groups who have pre-term scores that are located near this intersection point could be simply due to noise or random variation. That is, the post-term scores of the students in the two groups are statistically significantly different only for pre-term scores that are located some distance above and below that intersection point. The calculation the Johnson-Neyman confidence limits in Step 3 allows the researcher or program evaluator to determine the pre-term scores at which the post-term scores of the two groups are statistically significantly different. This information may lead the researchers or program evaluators to modify conclusions that were based solely on information provided by the analytical techniques contained in the first two steps of this process.

As was demonstrated by the analyses of the personal efficacy and teaching efficacy scores that were presented in this paper, following the three-step analytical procedure can provide essential information not only regarding whether an interaction effect does, in fact, exist but also with respect to the nature of the interaction effect. Such information can be invaluable to educational researchers and program evaluators.

References

Chou, T., & Huberty, C. J. (April, 1992). *The robustness of the Johnson-Neyman Technique.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Chou, T., & Wang, L. (April, 1992). *Making simultaneous inferences using Johnson- Neyman technique.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Fraas, J. W., & Newman, I. (April, 1977). *Malpractice of the interpretation of statistical analysis.* Paper presented at the annual meeting of The Ohio Academy of Science, Columbus, OH.

Gibson, S., & Dembo, M.H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology, 76,* 569-682.

Johnson, P.O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. In J. Neyman and E.S. Pearson (Eds.), *Statistical Research Memoirs,* 1936, 1, 57-93.

McNeil, K., Newman, I., & Kelly F. J. (1996). *Testing research hypotheses with the general linear model.* Carbondale, IL: Southern Illinois University Press.

Neter, J., Wasserman, W., & Kutner, M.H. (1985). *Applied linear statistical models* (2nd ed.). Homewood, IL: Irwin.

Newman, I., Deitchman, R., Burkholder, J., Sanders, R., & Ervin, L. (1976). Type VI error: Inconsistency between the statistical procedure and the research question. *Multiple Linear Regression Viewpoints,* 6 (4), 1-19.

Newman, I., & Fraas, J. W. (1979). Some applied research concerns using multiple linear regression [Monograph]. *Multiple Linear Regression Viewpoints,* 9, (4).

Newman, I., & Fry, W. (1972). A response to 'A note on multiple comparisons and a comment on shrinkage'. *Multiple Linear Regression Viewpoints, 2,* (1), 36-39.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction.* (2nd ed.). Fort Worth, TX: Harcourt Brace Jovanovich.

Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin, 88,* 307-321.

Rogosa, D. (1981). On the relationship between the Johnson-Neyman region of significance and statistical tests of parallel within-group regressions. *Educational and Psychological Measurement, 41,* 73-84.

Ross, J.A. (1994, June). *Beliefs that make a difference: The origins and impacts of teacher efficacy*. Paper presented at the meeting of the Canadian Association for Curriculum Studies, Calgary, Canada.

Russell, G. (1992). *FOCUS: An explanation of the human behavioral system.* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chicago, IL.

SPSS, Inc. (1990). SPSS/PC+ Statistics™ (Version 4.0) [Computer software]. Chicago: SPSS Inc.

SPSS, Inc. (1996). SPSS® Base 7.0 for Windows™ [Computer software]. Chicago: SPSS Inc.

Appendix

Computer Program for the Calculation of the Johnson-Neyman Confidence Limits.

```
Data list free/
  ss1  ss2  n1  n2  sumresid  mean1  mean2  slope1  slope2  int1
int2  fcrit
Begin data.
1434.21  1821.59  29  39  2495.58  39.31  38.50  .85  .31  6.36
31.49  3.99
End data.
Compute term1 = (fcrit/(n1+n2-4))*sumresid.
Compute terma = term1*(-1).
Compute a = ((terma)*((1/ss1)+(1/ss2)))+(slope1-slope2)**2.
Compute  b = (term1*((mean1/ss1)+(mean2/ss2)))+((int1-
int2)*(slope1-slope2)).
Compute c = (terma)*(((n1+n2)/(n1*n2))+((mean1**2)/ss1)+
                    ((mean2**2)/ss2))+((int1-int2)**2).
Compute RegionU = ((b*(-1))+(sqrt((b**2)-(a*c))))/a.
Compute RegionL =  ((b*(-1))-(sqrt((b**2)-(a*c))))/a.
List RegionU RegionL.
```

Using the World Wide Web: Suggested Applications and Precautions for Teaching Multiple Linear Regression

Lynne M. Pachnowski and **Isadore Newman**
University of Akron

The World Wide Web provides an excellent resource of real data that can be used by research methods instructors. Using this data in class allows more time for students to spend learning how to write appropriate statistical equations. This paper provides two ezamples of multiple linear regression equations that were written based on data available on the World Wide Web. It also provides a number of Internet references for instructors that would be helpful for obtaining real data that can be applied in a statistics course.

One of the most important processes a research methods instructor can help his/her students acquire is the ability to write a good research question. It is always beneficial for the instructor to have real data to use in formulating a question, since these questions are more likely to be perceived by the students as practical and applicable. The World Wide Web acts as a wonderful source of real data that instructors and students may access in a quick and convenient manner.

Once an instructor and the students have identified research questions using this data and the instructor feels that the students are competent in formulating questions, then the next step is to apply appropriate statistical models to test these questions. Implementing this instructional technique reduces the chance of committing a Type VI error which occurs when there is inconsistency between the research question being asked and the research model used to test it. (Newman & Newman, 1994). When the World Wide Web data is so readily available for demonstration, the instructor and students are able to spend far more class time discussing the analytical questions of which statistical models are appropriate to apply to the research questions developed. Therefore, the students benefit since they are more likely to be engaged in higher-order problem-solving activities.

The following are examples of research questions derived from World Wide Web data and multiple linear regression equations that can be applied to them.

Examples of Sites and Their Multiple Regression Applications

• The Wilmington Institute: Trial and Settlement Sciences

http://www.wilmington-institute.com/

The home page of this web site states that the institute was established in order to "help trial lawyers, corporate counsels and governmental agencies forecast the probable outcome of their litigation and trials." By choosing "Jury Talk Survey", the user is able to select a number of on-line survey of recent, well-publicized court cases, such as the O.J. Simpson case or the Timothy McVeigh case. The survey inquires what the visitor's opinion is regarding the nature of the case and requests some demographic information about the visitor. For instance the McVeigh questions are: "Do you believe Timothy McVeigh was involved in the planning and/or execution of the Oklahoma City federal building bombing? and If you answered yes to (1), do you believe Timothy McVeigh was part of a well organized and financed, geographically dispersed anti-government conspiracy?" The participant is then asked to identify his/her age group, gender, ethnicity, and area of residence from a list of possibilities. Once the results are submitted, updated overall results in terms of percentages appear on the screen.

With a data source such as this site, students could look at the results of a survey (guilty, not guilty, etc.) and the demographic information of the respondents to the survey and be encouraged to write sample research questions. For example, are there differences in age, ethnicity, gender, and geographic region and the verdict given? Such a question and the model for it might look like the following:

Is there a profile using age, ethnicity (Black, Hispanic, Asian, and other), gender and geographic region (Georeg1, ... , Georeg5) that differentiates those respondents who respond "guilty" and those who respond "not guilty"?

Model 1: $y = a_0u + a_1Age + a_2Sex + a_3Black + a_4Hisp + a_5Asian + other + a_7Georeg1 + a_8Georeg2 + a_9Georeg3 + a_{10}Georeg4 + a_{11}Georeg5 + a_{12}E1$
Model 99: $y = a_0u + E2$

All the independent variables in Model 1 are binary-coded (1, 0).

Looking at the same data, another question that could be raised could be whether ethnicity accounts for a significant amount of variance in their perception of guilt/no guilt over and above age, gender, and geographic region. This would be done by testing Model 1 against Model 2 given below:

Model 2: $y = a_0u + a_1Age + a_2Sex + a_3Georeg1 + a_4Georeg2 + a_5Georeg3 + a_6Georeg4 + a_7Georeg5 + aE3$

One can also test for interaction between pairs of variables such as age and ethnicity, age and gender, or gender and ethnicity since examples of addition models could not be tested.

• U.S. Census Bureau
http://www.census.gov

This site is provided by the U.S. Department of Commerce, Bureau of the Census. It contains a wealth of U.S. statistics, state and county statistics, and links to international census-related databases. The home page contains, among other items, links to the current U.S. and world approximated population counts, current economic indicators, and census documents.

A powerful link for researchers is the "International Data Base", which can be obtained by clicking through the path: "Current U.S. Population Count", "World", and finally "International Data Base". The visitor is then offered three links which offer three different ways of accessing the data provided. The visitor may either look at the data on the screen, load the data on a spreadsheet, or choose to configure appearance of the output. Once the output manner is selected, the user is then asked to select a statistically-related table (for instance, "life table values, by sex"), to select one or more countries from a table, and to select one or more years or to accept the latest available year as a default. The database will return the requested data or a message stating that the data was not available if it applies.

In one instance, we obtained a table containing the population of Canada by ethnic group and sex and also the U.S. population by ethnic group and sex. (After requesting the latest available year for each, the Canadian data provided was from 1991 and the U.S. data was from 1980. Since significant population changes probably occurred during those years, another search may want to be done to obtain 1980 data from Canada or similar data from another country from a year closer to 1991.) Data such as this would be helpful in teaching students how to write regression equations to test Chi Squares. A question derived from the data may be:

Proportionally, are there more men than women in the U.S. or Canada?

Model 3 $y = a_0u + a_1Males + E4$
Model 99a $y = a_0u + E5$

In the models above, y = 1 if male, and 0 otherwise. Also 1 = "from Canada" and 0 otherwise. By testing Model 3 against Model 99a, we can see if there is a significant difference in the proportion of males and females in the U.S. and Canada. The student would have to take the data given from the screen or spreadsheet and recode it in a manner ("zeroes" and "ones") that would be most effective for the statistical analysis. Since the data are presented in aggregated form, the student will have to learn how to put in the data in individual form from the aggregate. Also, the student will consequently learn how handle data that is provided in different formats. This could be part of the instructional process.

The above examples are only two of the many World Wide Web sites that can be used to obtain data that could be used to make examples more realistic. Many other sites are also available with similar or even more extensive data than the sites mentioned. One excellent archive of links to "data and depiction's of data from throughout the world" is "Dr. B's Wide World of Web Data" found at http://seamonkey.ed.asu.edu/~behrens/siip/webdata (then choose "Wide World of Web Data) and created by Dr. John Behrens of Arizona State University. This site contains thirty-one links to sites that contain data sets or data-related information. The data sets are separated into sixteen categories, including "Children and Youth", "Demographics", "Education", and "Social Science -- General". The page encourages instructors to use the data for examples in class and to encourage students to find data that they find interesting.

Instructional Precautions

An instructor that chooses to integrate information of the World Wide Web into a course needs to be aware of the advantages and disadvantages of using such a medium. Because of all the attention the Internet has received in the media, the advantages -- accessibility of data, student convenience -- may seem more evident than the disadvantages. Instructors and students both must be aware of some of the following cautions:

• Each server hosting a web site only has the capability of hosting a finite number of users. Therefore, some sites may not have the capability of hosting a class of twenty students each attempting to visit the site at one time. Therefore, in-class lab time experiences should be planned so that students have a variety of sites from which to choose. If a visit to a particular site is required, instructors should assume that the average student may need to make two to three attempts on different occasions in order to make a connection.

• While new sites are appearing on the Web every day, old ones are often neglected. Sometimes, a promising link title or URL address may, in fact, have no file at the end of it. Furthermore, many sites containing data may left to become obsolete. A user should look for a notation on the page as to when the page was last updated.

• One of the most important cautions to students and instructors of research is the format that "data" can take on the Web. Although the census site has data sets that may be downloaded and the Wilmington site is both interactive and provides overall results, some sites have a much more limited interface. A site that is promoted as a "database" may only have a search engine interface which keeps the entire database hidden from the user. A user may only view pieces of the database based on the parameters of the search he/she submits.

Although these precautions may seem daunting to some new users, a reasonably proficient Web user can address these precautions by simply testing each Web source before a classroom application and creatively designing classroom assignments involving the Web.

Despite the precautions of using the Web within a statistics course, it is still difficult to deny the long-term advantages of using the Internet data as both a teaching and research tool. Among these advantages is that students are able to gain experiences in working with a medium that is increasingly more likely to be a primary source of data in the student's home and workplace.

A paper presented at the 1997 Eastern Educational Research Association's National Meeting related to this topic can be found at:
http://junior.apk.net/~jurczyk/eera.html.
The paper contains links to the sites mentioned above as well as the following other related sites:

Government Resources:

U.S. Census Bureau
http://www.census.gov
Census reports and links to other federal government and international agencies offering statistical reports.

Fedworld
http://www.fedworld.gov
Central location and starting point for finding U.S. government information.

Government Statistics on the Internet (paper)
http://www.stats.gov.nt.ca/Bureau/General/WWWPaper.html
Survey of government statistics (Canada, U.S., U.K.) available on the Internet.

SEC (Securities and Exchange Commission)
http://www.sec.gov
U.S. government site includes filings by public companies.

Stat-USA
http://www.stats-usa.gov
Department of Commerce service offering detailed government statistics-based reports.

Other Resources:

Facts on File
http://www.facts.com
Producer of comprehensive studies of modern issues. Reports include some survey results with statistics.

The Gallup Organization
http://www.gallup.com
Provider of public opinion poll data.

The Harris Poll
http://techsetter.com/harris/html/home.html
Contains the latest Harris poll and comparisons of the previous poll's telephone responses with Internet responses.

CollegeNet
http://www.collegenet.com/
A directory of colleges and universities divided into various categories and search parameters.

Texas Lotto
http://crashdummy.iglobal.net/lotto
The results of the latest Texas Lotto drawing and the results of the drawing over several years.

References

Braun, E. (1994). *The Internet Directory*. New York: Ballantine Books.

Ellsworth, Jill H. (1994). *Education on the Internet: A Hands-On Book of Ideas, Resources, Projects, and Advice*. Indianapolis, IN: Sams Publishing.

Hahn, H. and Stout, R. (1994). *The Internet Yellow Pages.* New York: Osborne McGraw-Hill.

Newman, I. and Newman, C. (1994). *Conceptual Statistics for Beginners.* Lanham, Maryland: University Press of America.

Place, R. Dimmler, K. Powell, T. (1996). *Educator's Internet Yellow Pages.* Englewood Cliffs, N.J.: Prentice-Hall.

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.

$ITBS\_GAIN\_R96\_R95_{ij}$ = 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:

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

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

$E[u_{0j}] = 0$, $Var[u_{0j}] = \tau^2$, 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:

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

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:

$$\begin{aligned} ITBS\_RES\_R\_96_{ij} = {} & \beta_0 + \\ & \beta_1 ITBS\_RES\_R\_95_{ij} + \\ & \beta_2 ITBS\_RES\_M\_95_{ij} + \varepsilon_{ij} \end{aligned}$$

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

Model 4

STAGE 1:

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

STAGE 2:

Level 1:

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

where

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

Level 2:

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

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

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

Model 5

STAGE 1:

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

STAGE 2:

Level 1:

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

Level 2:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{kj} &= \gamma_{k0} \qquad \text{for } k = 1, 2. \end{aligned}$$

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

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

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 $k = 0, 1, 2$.

$E[u_{kj}] = 0$, 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 $k = 1, 2$.

$E[u_{0j}] = 0$, 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

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

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

Model 10

Level 1:

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

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:

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

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:

$$ITBS\_RES\_R\_96_{ij} = \beta_{0j} + \beta_{1j} ITBS\_RES\_R\_95_{ij} + \beta_{2j} ITBS\_RES\_M\_95_{ij} + \beta_{3j} ITBS\_RES\_R\_94_{ij} + \beta_{4j} ITBS\_RES\_M\_94_{ij} + \beta_{5j} ITBS\_RES\_R\_93_{ij} + \beta_{6j} ITBS\_RES\_M\_93_{ij} + \delta_{ij}$$

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:

$$ITBS\_GAIN\_R95\_R96_{ijk} = \pi_{0jk} + \pi_{1jk} ITBS\_R\_95_{ijk} + \pi_{2jk} ITBS\_M\_95_{ijk} + \pi_{3jk} ITBS\_R\_94_{ijk} + \pi_{4jk} ITBS\_M\_94_{ijk} + \pi_{5jk} ITBS\_R\_93_{ijk} + \pi_{6jk} ITBS\_M\_93_{ijk} + \varepsilon_{ijk}$$

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

Level 2:

$$\pi_{pjk} = \beta_{p0k} + \beta_{p1k} BLACK_{jk} + \beta_{p2k} HISPANIC_{jk} + \beta_{p3k} GENDER_{jk} + r_{pjk}$$

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

Level 3:

$$\beta_{p0k} = \gamma_{00k} + u_{p0k}$$
$$\beta_{pqk} = \gamma_{p0k} \quad \text{for } q = 1, 2, 3$$

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

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

Model 14

STAGE 1:

$$Y_{ij} = \Lambda_0 + \Lambda_1 BLACK_{ij} + \Lambda_2 HISPANIC_{ij} + \Lambda_3 GEND\ R_{ij} + \varepsilon_{ij}$$

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:

$$ITBS\_RES\_R\_96_{ij} = \beta_{0j} + \beta_{1j} ITBS\_RES\_R\_95_{ij} + \beta_{2j} ITBS\_RES\_M\_95_{ij} + \beta_{3j} ITBS\_R\ S\_R\_94_{ij} + \beta_{4j} ITBS\_RES\_M\_94_{ij}$$

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

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.)

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

The shrinkage coefficient is

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

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:

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

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1}T_{1j} + \gamma_{k2}T_{2j} + \ldots + \gamma_{k10}T_{10j} \quad \text{for } k = 1, 2.$$

$E[u_{0j}] = 0$, $Var[u_{0j}] = \tau^2$, 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:

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

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1}T_{1j} + \gamma_{k2}T_{2j} + \ldots + \gamma_{k10}T_{10j} \quad \text{for } k = 1, 2.$$

$E[u_{0j}] = 0$, $Var[u_{0j}] = \tau^2$, 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:

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1}T_{1j} + \gamma_{k2}T_{2j} + \ldots + \gamma_{k10}T_{10j} + u_{kj} \quad \text{for } k = 0, 1, 2.$$

$E[u_{0j}] = 0$, $Var[u_{0j}] = \tau^2$, 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:

$$\beta_{00k} = \gamma_{000} + u_{00k}$$
$$\beta_{pqk} = \gamma_{pq0} \quad \text{for all other } p \text{ and } q.$$

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

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:

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

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

Table 3 Correlations of The School Effectivenss Indices with Important Student Contextual Variables, Grade 6

Table 4 Correlations of The School Effectivenss Indices with Important Student Contextual Variables, Grade 8

Table 5 Correlations of The School Effectivenss Indices with Important School Contextual Variables, Grade 6

Table 6 Correlations of the School Effectivenss Indices with Important School Contextual Variables, 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.

MINUTES

OF THE

ANNUAL MEETING

OF THE

MULTIPLE LINEAR REGRESSION: GENERAL LINEAR MODEL / SIG

(Chicago, IL)

MARCH 27, 1997

Professor Randy Schumacker (University of North Texas), SIG Chair, opened the business meeting. The first order of business was approval of the April 11, 1996 SIG MLR:GLM meeting minutes as distributed in the Multiple Linear Regression Viewpoints (Spaner, MLRV, 23(1), p. 35). No corrections or changes were offered and the minutes were approved as distributed.

Schumacker reported on the success of the SIGs sessions at this year's AERA conference: the HLM session attendance was 117 and the GLM session attendance was 25. Schumacker suggested that the SIG sponsor a theme session at the 1998 AERA conference on longitudinal analysis. Schumacker also indicated that he intended to vigorously pursue this year's presenters at the SIG sessions to urge their submission of their papers to the MLRV for publication. Schumacker volunteered to prepare a SIG MLR:GLM newsletter for distribution before the 1998 AERA conference outlining the SIG's conference activities.

The Chair then called upon the SIG Executive Secretary, Steve Spaner, to give his report. Spaner presented the budget report that was being submitted to AERA headquarters for the 1996-97 year. The SIG treasury was $2263.07 on 4-1-96, the SIG account has earned $51.83 interest over the year and received $925.55 in member dues for a total assets of $3240.45 on 3-1-97. The SIG incurred expenses of $227.43 since 4-1-96 leaving the SIG with a $3013.02 balance on 3-1-97. Spaner reported that the current paid membership was N=70 on 3-1-97. Spaner indicated that the SIG decline in membership has been correlated with the reduced number of issues of and irregular schedule for the Multiple Linear Regression Viewpoints (MLRV), the MLR:GLM/SIG's journal. Journal editor John Pohlmann urged members to submit articles and comments for consideration in MLRV. It was suggested, once again, that persons making presentations under the MLR:GLM/ SIG sponsorship at the AERA conference should be required to submit their papers to the MLRV. No motion to that effect was made. Schumacker informed the members that the AERA SIG Committee had announced a new formula for the assignment of SIG sessions: one session for every 43 AERA members. Schumacker stressed the importance of not only recruiting SIG members but members who also held AERA membership.

Schumacker moved to the New Business part of the meeting and election of officers. Executive Secretary, Steve Spaner (University of Missouri - St. Louis), explained that the MLR:GLM/SIG election procedures call for the election to be held by mail ballot and the business meeting to be a nominating meeting only. As there were no prior nominations in response to the call in the MLRV (1997, 23(1)), the chair opened the floor for nominations. It was moved and passed by the members attending to suspend the election by mail ballot rule and to hold the election at the business meeting. The first call was for Chair-elect nominations. Isadore Newman (The University of Akron) was the sole nomination. Therefore, it was moved and seconded that Dr.

Newman be elected by acclamation. Motion passed. The next set of positions to be elected were two replacement Executive Board/Editorial Board members. Again, as there were no prior nominations in response to the call in the MLRV (1997, 23(1)), the chair opened the floor for nominations. The nominated Executive Board/Editorial Board replacements were Professor John Dixon (University of Florida) and Dr. Werner Wothke (SmallWaters Corporation, Chicago). With no additional nominations made, it was moved and seconded that Professor Dixon and Dr. Wothke be elected by acclamation. Motion passed. Drs. Dixon and Wothke replace Drs. Gregory Marchan (Ball State University) and John Williams (University of North Dakota) and assume the four year terms from 1997-2001.

Respectfully submitted,

Steven D. Spaner,
Executive Secretary

SPECIAL NOTICE

TO: LIBRARIES AND INSTITUTIONS (and MLR:GLM/SIG members)

RE: VOLUMES 18 - 24 (1991 - 97) of Multiple Linear Regression Viewpoints

The EBSCO and FAXON subscription services have been notified in each of the years listed above that the MLR Viewpoints has reduced its publication frequency to "occasional." While we strive to put out two issues per year (i.e., two issues per volume), for the past seven years (7 volumes) we have had insufficient submissions to make a second volume economical. We still hold to our goal of two issues a year, but do not guarantee two issues per year and do not honor claims for a second issue (i.e., the succeeding years' issue) in years when no second issue was published. We hope this clears up a number of outstanding claims notices. We thank you for your support of and interest in our journal and our Special Interest Group.

Sincerely,

John Pohlmann, PhD
Editor, MLR Viewpoints
Department of Educational Psychology
Southern Illinois University-Carbondale
Carbondale, IL 62901
e-mail: johnp@siu.edu

Steven Spaner, PhD
MLR:GLM/SIG Executive Secretary
Division of Educational Psychology
University of Missouri-St. Louis
St. Louis, MO 63121-4499
e-mail: sspaner@umslvma.umsl.edu

(Secretary's note: 1998 membership payment is due at the beginning of the 1998 calendar year. If the first line of your mailing label ends in 97, you now owe for the 1998 MLR:GLM/ SIG membership year. If your mailing label has 96 or earlier at the end of the first line, you are unpaid for the past 1997 MLR:GLM/SIG membership year as well as owe 1998 MLR:GLM/SIG membership dues)

MULTIPLE LINEAR REGRESSION: THE GENERAL LINEAR MODEL SPECIAL INTEREST GROUP

APPLICATION FORM FOR NEW MEMBERS / RENEWAL MEMBERS

Membership in the Multiple Linear Regression: the General Linear Model Special Interest Group (MLR:GLM/SIG) of AERA entitles the member to participate in all the activities of the MLR:GLM/SIG. These activities include: participation in the MLR:GLM/ SIG annual meeting and social gathering, voting for MLR:GLM/SIG officers and Multiple Linear Regression Viewpoints (MLRV) editors, and the right to contribute articles and papers to the MLRV (a periodic publication of referred articles, invited topic papers, issues and news items, and other communications to the MLR:GLM/SIG membership). Membership includes a 1 year subscription to the MLRV.

There are three forms of membership in the MLR:GLM/SIG: individual, student and library/institutional: **Individual membership dues are $10.00 per year** (or $18 for two years), student membership dues are $5.00 per year, and library/ institutional fees are $20.00 per year. Dues and fees for each year are due January 1 of the year, but, payable NO LATER THAN the annual business meeting of the MLR:GLM/SIG. The business meeting is held during the annual AERA Convention which meets within the week prior to or following Easter Sunday of each year.

COMPLETE this bottom SECTION and SEND it with your PAYMENT.

PLEASE PRINT:

NAME:_____

MAILING ADDRESS _____

_____

_____

city state

_____ _____

post code (zip) country

E-MAIL ADDRESS _____

MEMBERSHIP TYPE:
(Check one in each set)

Individual: 1 Year ($10)____
2 Years($18)____

Student: ($ 5)____

Library: ($20)____

New Member ____

Renewal Member ____

Member of AERA?
Yes ____ No ____

Membership dues and MLRV subscription fees should be made payable to:
Multiple Linear Regression SIG

and sent to: Steven Spaner [e-mail: sspaner@umslvma.umsl.edu]
MLR:GLM/SIG Executive Secretary
Division of Educational Psychology
UM-St. Louis, 8001 Natural Bridge Road
St. Louis, MO 63121-4499 USA

(10Jan98)

MULTIPLE LINEAR REGRESSION: THE GENERAL LINEAR MODEL SPECIAL INTEREST GROUP

OFFICER NOMINATION FORM

Chair-elect (1999 Chair)
[one year term]
Name(s) and Affiliation(s): _____

Executive Secretary (1998-2001)
[three year term]
Name(s) and Affiliation(s): _____

MLRV Editor (1998- TDOR)
[term open]
Name(s) and Affiliation(s): _____

Two (2) Editorial/Executive Board Members (1998-2002)
[six year term]
Name(s) and Affiliation(s): _____

Name(s) and Affiliation(s): _____

==========

Please send form to (due by April 3, 1998):

Steven Spaner [e-mail: sspaner@umslvma.umsl.edu]
MLR:GLM/SIG Executive Secretary
Division of Educational Psychology
UM-St. Louis, 8001 Natural Bridge Road
St. Louis, MO 63121-4499 USA

Or fax to 314-516-5784

(10Jan98)