DOCUMENT RESUME

ED 346 570 EA 023 996

AUTHOR Bernstein, Lawrence
TITLE The Development of a Multilevel Model of State Level Student Achievement. Pennsylvania Educational Policy Studies Number 5.
INSTITUTION Pittsburgh Univ., Pa. Learning Research and Development Center.; Pittsburgh Univ., Pa. School of Education.
PUB DATE Oct 90
NOTE 28p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Data Analysis; Grade 5; Intermediate Grades; *Multiple Regression Analysis; *Predictive Measurement; *Regression (Statistics); Research Methodology; Research Problems; *State Norms
IDENTIFIERS *Hierarchical Linear Modeling; *Pennsylvania

ABSTRACT

Educational research on the factors of student achievement has been limited by its failure to consider the multilevel or hierarchical nature of most data. This study used a nonexperimental regression-based procedure, hierarchical linear modeling (HLM), to empirically develop a predictive model of fifth-grade achievement in reading and mathematics for a statewide data set at both the individual student and school district levels. The database was comprised of reading and mathematics achievement test scores of 86,227 elementary students in Pennsylvania who were enrolled in third grade in 1986 and in fifth grade in 1988. Findings indicate that only a small portion of the variability in individual achievement is potentially explainable by district-level factors. HLM was also used to identify district-level factors that explain the variation in district mean achievement and within-district relationships. For example, a small effect of class size was revealed in increasing the within-district relationship between prior ability and student achievement. These results permit the formulation of a wider range of policy inferences than is possible with conventional regression analyses. One figure and six tables are included. (27 references) (LMI)

Pennsylvania Educational Policy Studies

PEPS is a joint effort of the U. of Pittsburgh's School of Education and the Learning Research and Development Center
This is policy paper number 5 in this series

The Development of a Multilevel Model of State Level Student Achievement

by

Lawrence Bernstein

University of Pittsburgh

October 2, 1990

The purpose of this series of papers is to contribute to a more informed debate about critical policy issues facing Pennsylvania's public schools. This PEPS series draws upon a data base that has been established here at the University of Pittsburgh under the direction of William Cooley in cooperation with the Pennsylvania Department of Education.

Reactions can be shared:
by mail: LRDC, Pgh.,PA 15260
by PittVAX: LARRYB
by phone: 412-648-8805
by BITNET:LARRYB@PITTVMS

Abstract

A continuing concern of educational researchers has been determining the factors that contribute to promoting student achievement. Those research efforts that were primarily focused at the district or school level have been remiss in taking into account the multilevel or hierarchical nature of most educational data. For educational research to have policy relevance, the methodological mismatch between the multilevel nature of educational data and the use of linear unilevel data-analytic models needs to be resolved.

This study used a nonexperimental regression-based procedure, hierarchical linear modeling (HLM), to empirically develop a predictive model of fifth-grade achievement in reading and mathematics for a large state-wide data set at both the individual student and school district level. The results showed that only a small portion of the variability in individual student achievement is potentially explainable by district-level factors. In addition, HLM was used to identify district-level factors which explain not only the variation in district mean achievement, but within-district structural relationships as well. The HLM analysis revealed, for example, a small effect of class size in increasing the within-district relationship between prior ability and student achievement. These results permit a wider range of policy inferences to be made than would be possible with conventional regression procedures.

Acknowledgements

This study draws upon a data base that has been established by Pennsylvania Educational Policy Studies (PEPS) under the direction of William Cooley. PEPS is a joint effort of the University of Pittsburgh's School of Education and the Learning Research and Development Center acting in cooperation with the Pennsylvania Department of Education.

Introduction

The report *Equality of Educational Opportunity* (Coleman et al., 1966) questioned the effect of schooling on student achievement and spawned numerous studies (see Glasman & Biniaminov, 1981; Madaus, Airasian & Kellaghan, 1980; and Mosteller & Moynihan, 1972 for examples) aimed at investigating the relationship between a set of input variables and student achievement within a given school system. The studies comprising the "school effects" literature have often arrived at conflicting conclusions due to the use of divergent sampling procedures, different units of analysis, disparate data analysis techniques, and varying operational definitions of the input variables and/or the outcome measures.

One of the most glaring problems has been the failure to take into account the multilevel nature of the data inherent in most educational settings. In other words, variables of interest are often observed and measured at different levels of analysis, e.g., individual students and school districts. For the field of educational research, this has resulted in the absence of a consensus as to what factors promote student achievement. Moreover, from an educational policy perspective, this lack of consensus has led to inconsistent and counterproductive applications of empirical inquiry revolving around a number of policy questions such as school district consolidation or equity in school financing (see Geske, 1983; Guthrie, 1979).

Recent studies (e.g., Bidwell & Kasarda, 1975; Friedkin & Necochea, 1988; Turner, Camilli, Kroc & Hoover, 1986; Walberg & Fowler, 1987) have investigated the relationship between so-called input variables, notably socioeconomic status, district size and expenditures, with student achievement at the district level within a given state. The results of these studies suggest that the role of input variables in explaining variation in student achievement is largely a function of the level of data aggregation (see Blalock, 1964; Hannan, 1971; Langbein, 1977 for further clarification).

Thus, proper specification of a predictive model of student achievement entails not only the inclusion of relevant predictor variables related to achievement (Cooley, 1978), but an incorporation of the multilevel nature of the data as well. In other words, developing an adequate predictive model of student achievement requires that the multilevel nature of a state educational system be taken into account by indicating how variables at one level of the system (e.g., district) might interact with variables at another level (e.g., student).

For educational research to have policy relevance the methodological mismatch between the hierarchical (i.e., nested/multilevel) nature of most educational data and the use of linear unilevel data-analytic models needs to be resolved. Previous studies of this scope and magnitude have ignored the multilevel nature of educational data, and consequently have drawn causal inferences which have obscured rather than clarified specific policy issues. Cronbach (1976) was one of the first investigators of the multilevel phenomenon to put it bluntly:

> The majority of studies of educational effects -- whether classroom experiments, or evaluations of programs, or surveys -- have collected and analyzed data in ways that conceal more than they reveal. The established methods have generated false conclusions in many studies (p. 1).

Failure to take the multilevel nature of the data into account can lead to incomplete or incorrect empirical inferences with undesirable consequences for informing state-level policy questions. In the past, the use of available aggregate-level data has been justified on the grounds that collecting individual-level information was too costly (Langbein, 1977). But as Aitkin and Longford (1986) ask:

> If the analysis of available aggregate data leads to wrong conclusions and disastrous educational policies, where is the economic or educational benefit to be found? (p. 42).

Questions of research design also have consequences regarding the issue of multilevel data. In investigating the impact of particular variables within a policy context, e.g., the effect of school district size or educational expenditures on student achievement, it is unfeasible and unrealistic to utilize experimental design methodology requiring random assignment. For example, students cannot be randomly assigned to school districts of varying size or rate of educational expenditures. Thus, educational evaluation or decision-oriented research efforts often must rely on naturally occurring data to investigate the role of a group of input variables in explaining variation in student achievement (Cooley, 1978; Longford, 1989). Unfortunately, the multilevel nature of educational data is not adequately addressed with conventional methodologies (Burstein, 1980; Cooley, Bond & Mao, 1981; Cronbach, 1976).

Purposes of Research

This paper is based on a broader study (Bernstein, 1990), which has two overlapping purposes. One is to develop and test a data-based predictive model which identifies the relevant variables associated with student achievement from a multilevel perspective, i.e., students within districts. The model is developed from an extant data base using conventional ordinary least squares regression techniques and the regression-based hierarchical linear modeling procedure, HLM (Bryk et al., 1988). The second purpose is to examine the differences in which parameters are estimated in a multilevel analysis using HLM in comparison to conventional analyses conducted at a single level of measurement (individual student or school district).

This paper addresses in part these purposes by discussing the limitations of using linear unilevel regression procedures in modeling student achievement. An introduction to the hierarchical linear modeling procedure is also provided as an illustration of the potential application of multilevel analysis in providing information to policy makers at the state level.

Data Base

The data base for this study is comprised of the cohort of Pennsylvania elementary school pupils who were enrolled in third grade in 1986 and in fifth grade in 1988, and who participated in the statewide Tests of Essential Learning Skills (TELS) program in reading and mathematics. The analyses described in this paper are based on a total population of 86,227 students enrolled in 1794 elementary schools (in 1988) within a total of 496 school districts throughout the state of Pennsylvania. The outcome measure for this study is a composite academic achievement score (ACH88) comprised of a weighted average of the fifth grade reading and mathematics test scores. In one sense, the tests are considered to be criterion-referenced measures since they indicate whether students have met state-mandated minimum competency levels in basic learning skills in reading and mathematics In this capacity, the tests employ cut-off scores to identify students in need of remedial instruction.

In addition to the outcome measure, several other variables are measured at the student level. One key variable is a measure of prior student ability, a composite achievement score (ACH86) based on performance on the third grade TELS reading and math measures. Other student background variables that are obtained are SEX (1=male/2=female), RACE (1=nonwhite/2=white), special education status (SEC88), compensatory education participation in reading and math (CEPRM88), and a home environment-motivation index (HOMEMOTV).

Variables at the district level are grouped into several categories. Three district-level proxies for socioeconomic status are included: a measure of a market value/personal income aid ratio for 1988 (AIDRATIO); the percentage of families receiving federally-sponsored assistance for dependent children (AFDCPCT) and an aggregated mean of the home environment/motivation index (AVGHOMTV). In addition, a measure of average daily attendance divided by enrollment (ATTEND) is included to tap the level of motivation in a school district.

School district size is represented by an enrollment variable (LNADM) measuring average daily membership over the school year 1987-88. A logarithmic transformation is used to correct for nonlinearity of district size effect estimates due to the presence of a few extremely large school districts. An additional size variable (POPMILE) measures population density per square mile to reflect an urban vs. rural distinction among school districts.

Expenditures for schooling is measured by a number of 1987-1988 fiscal indicators. Instructional cost per pupil (EXPEND) is derived by dividing the net instructional expenditures of a district by average daily membership. Variables related to expenditures for schooling at the district level are represented by average district class size (CLSIZ), mean elementary school teacher salary in the district (TCHSAL), and ratio of the number of students to the number of district administrative personnel (ADMIN). In addition, variables measuring teacher experience (TCHSERV) and level of training (TCHLEVEL) are included.

Finally, a number of aggregated measures are included to examine the degree of contextual effects. Contextual effects are defined here as the influence of peer group composition on student achievement. PINSE88 measures the effect of the percentage of special education students in the district. PCTNW is the percentage of nonwhite (black and Hispanic) students in the district. DCEPRM88 is an aggregated measure of compensatory educational status at the individual student level. An additional contextual effect (DCUTTELS) measures the aggregated mean score of students falling below the cutoff on either reading or mathematics in a school district.

Data Analytic Procedures

The empirical development of a predictive model of student achievement was based on a series of individual student and district-level ordinary least squares regression analyses. In the district-level analysis the outcome measure (DACH88) was created as an aggregated mean of district-level achievement and regressed on a set of district-level indicators and aggregate measures of the student background variables. An individual level regression analysis was conducted with the composite achievement score (ACH88) and the student-level predictor variables.

Informed by the above results, a set of analyses were then performed using the hierarchical linear modeling program, HLM. In its simplest form, a hierarchical linear model is comprised of two equations describing the structural relationships in a within-and a between-group model (see Raudenbush and Bryk, 1986, 1988). The regression coefficients estimated in the within-group model, β_{jk}, (for k predictor variables in j groups) constitute the "microparameters" in the HLM formulation, which are then modeled as the outcome variables in the between-group model. At this level in the hierarchy the β_{jk} regression coefficients are postulated to vary across groups as a function of "macroparameters", θ_{pk}, which represent the systematic effects of *p* group-level variables on the *k* within-group relationships. Of key interest in HLM is the capturing of the variation of the structural relationships across groups through the creation of both a within- and a between-group model. HLM uses information from both levels of the model (e.g., individual students and school districts) and therefore does not force researchers to choose between the two units of analysis.

Results of District- and Individual-Level Analyses

In the district level analysis, an initial set of 16 predictors with DACH88, average fifth grade achievement, as the outcome variable was entered into a multiple regression analysis for the purpose of determining the relative importance of the input variables.[1] In order to reach a more parsimonious, regression coefficients with t-ratios less than 1.5 were deleted until a model was reached with five predictors: DACH86, AVGHOMTV, AFDCPCT, LNADM, and CLSIZ. Table 1 presents the results from the district-level analysis. The zero-order Pearson correlation coefficient with DACH88 is represented by r, while *B* and Beta refer to the unstandardized and standardized regression coefficients respectively.

In this model which accounted for 51 percent of the observed variance in achievement, DACH86 appeared to be the strongest predictor in terms of explaining variance in district achievement.[2] AFDCPCT displayed a moderate negative relationship with achievement, with

---

[1]The model-building strategy employed here is primarily of an empirical nature to allow for comparisons with a multilevel analysis. The use of empirical procedures for variable selection in regression analysis has justifiably been criticized for its substantive short-sightedness (see Boyd & Iversen, 1979 for examples).

[2]It should be kept in mind, however, that regression coefficients are unreliable indicators of the strength of relationship between two variables.

Table 1
Regression Analysis for District-Level Model

Multiple R	.719
R Square	.518
Adjusted R Square	.513
Standard Error	2.536

VARIABLE	r	B	BETA	T	p
DACH86	.672	.487	.490	12.331	.000
AVGHOMTV	.513	1.849	.184	4.537	.000
AFDCPCT	-.458	-.083	-.175	-4.727	.000
LNADM	-.034	-.414	-.077	-2.329	.020
CLSIZ	-.102	-.386	-.109	-3.407	.001

LNADM and CLSIZ more weakly related.[3] AVGHOMTV exhibited a moderate positive relationship in accordance with its zero-order correlation with DACH88.

What is the effect of deleting DACH86 from the model? The regression coefficients for LNADM, AVGHOMTV, CLSIZ and AFDCPCT all increased in magnitude to varying degrees. Removing DACH86 from the model not only reduces the Adj. R^2 from .512 to .363, but affects the magnitude of the other parameters in the model as well.

For purposes of mparison, a comparable set of analyses were conducted on the individual student level variables and achievement. A multiple regression analysis was run with ACH88 as the outcome variable and six predictor variables, ACH86, CEPR88, RACE, SEC88, SEX and HOMEMOTV. This model explained about 64 percent of the observed variance in achievement. ACH86 was the most powerful predictor as indicated by its beta weight. SEC88 and CEPR88 had negative regression coefficients, while RACE and HOMEMOTV were positively related to ACH88. There was little difference in achievement due to the effect of SEX. On the basis of these results, SEX was dropped from the analysis and the individual level model was refitted with the remaining five predictors. This final model

[3]Both LNADM and CLSIZ demonstrated a "suppressor" effect in that their partial-order correlation coefficients are higher than their zero-order correlations with achievement. These variables were retained in the analysis, however, due to their policy relevance.

did not suffer any drop in the adjusted R^2 figure and the model estimates remained virtually identical. Table 2 presents the results for this model with correlation coefficients (r) and unstandardized (*B*) and standardized (Beta) regression weights.

As with the district level regression, an analysis was run without ACH86 to investigate the effect of deleting prior ability on model specification. An analysis with the remaining four predictor variables had an adjusted R^2 figure of .411, a decrease of about one-third from the previous model. In this model, all the effects increase dramatically in size and remain in the same direction.

Table 2
Regression Analysis for Individual-Level Model

Multiple R	.797
R Square	.636
Adjusted R Square	.636
Standard Error	8.987

VARIABLE	r	*B*	BETA	T	p
ACH86	.776	.675	.622	230.687	.000
HOMEMOTV	.349	.875	.082	37.224	.000
RACE	.342	5.037	.109	49.883	.000
CEPRM88	-.460	-3.462	-.122	-52.014	.000
SEC88	-.325	-6.854	-.095	-42.774	.000

Analyses conducted on the individual level can account for variables measured at a higher level of aggregation by disaggregating the effects back to the individual level (see Summers & Wolfe, 1977 for an example). In this manner, contextual effects can be measured by including a constant value associated with the school district for each individual student record. A model including the contextual effects of PCTNW, AVGHOMTV, DCEPRM88, DCUTTELS, and PINSE88 increased the explained variation in achievement by about two percentage points to an adjusted R^2 of .655. Of more importance, however, is the interpretation of the regression coefficients in this model. In this new model with district level contextual effects predicting student achievement, the individual level effects remained the same, except for RACE, which decreased by more than 50 percent. The coefficients for the contextual effects, PCTNW, PINSE88 and DCEPRM88 were all positive whereas their zero-order correlations with achievement were negative. AVGHOMTV, on the other hand, had a negative effect on achievement in this model as opposed to its positive zero-order correlation with ACH88. Only DCUTTELS had the expected effect: The higher the district mean score of students below the cutoff on reading and math, the lower an individual student's score with the effects of the other predictors in the model held constant.

The explanation for this counter-intuitive pattern of results probably lies in the simultaneous modeling of a group of individual and group-level effects. Not only are the individual variables correlated with their group-level counterparts, but the group-level variables are also intercorrelated with each other.[4] Multicollinearity within the context of simultaneously modeling individual and group-level effects creates a difficulty in interpreting the relative importance of the explanatory variables in the model (Boyd and Iversen, 1979).

The resolution of this problem lies in adopting a technique which allows for the explicit modeling of individual-level relationships as a function of group-level factors. The following section of the results is devoted to a review of the analyses employed with the hierarchical linear modeling (HLM) program.

Results of Multilevel Analysis

An HLM model was formulated for the purposes of analyzing district mean variability. In this analysis, information was initially provided as to how much variation in the outcome measure, ACH88, lay within and between districts. The model, equivalent to a one-way random effects analysis of variance with districts treated as a random factor, was posed as:

[4]The correlations among the district-level contextual effect variables at the individual level are higher than their corresponding intercorrelations at the district level.

$Y_{ij} = \mu_j + \varepsilon_{ij}$, (within-district) and

$\mu_j = \theta + U_j$ (between-district)

In this formulation, the within-district model states that the outcome variable Y_{ij} (fifth grade academic achievement of student *i* in district *j*) varies around a district mean μ_j with independent errors ε_{ij} assumed to be distributed ~ $N(0,\sigma^2)$ where σ^2 signifies within-district variance. In turn, in the between-district model, each district's mean μ_j varies around a grand mean θ with independent errors U_j assumed to be distributed ~ $N(0,\tau)$ with τ signifying between-district variance.

The ratio of $\hat{\tau}$ (estimated between-district variance) to $\hat{\tau} + \hat{\sigma}^2$ (estimated between- and within-district variance) yields an intra-district correlation coefficient $\hat{\rho}$, which expresses the estimated proportion of variance in the outcome measure between districts. In this study, $\hat{\tau} = 11.28$ and $\hat{\sigma}^2 = 185.09$, with $\hat{\rho} = .057$ indicating that approximately six percent of the variance in Y is located between districts.

The next question concerned whether district means varied significantly across districts, or H_0: $\tau = 0$, where τ again represents the amount of between-district variation in terms of means. In a large-sample test of this hypothesis, the Chi-square test statistic was equal to 19132.0 with 495 degrees of freedom, $p<.001$, indicating that the null hypothesis could be rejected and that districts did show significant variability in mean achievement.

The final question centered around determining the contribution of district-level factors to explaining the variability in district mean achievement. Note that if $\tau = 0$, then district-level factors cannot explain variability in district means. In this between-district model, each district mean score is predicted by district factors such as district mean home environment/motivation (SES):

$\mu_j = \theta_0 + \theta_1(\text{mean SES})_j + U_{0j}$,

where

θ_0 is equal to the grand mean of achievement,

θ_1 is equal to the effect of district mean SES on μ_j,

and

U_{0j} is assumed to be distributed ~ $N(0,\tau)$, where τ now represents the residual parameter variance after controlling for district mean SES.

The results showed a highly significant relationship between district SES as measured by home environment-motivation and district mean achievement (θ_1 = 5.12, t = 13.26, p<.001). The residual parameter variance, τ, after accounting for the district SES factor is now reduced to 7.90 from 11.28. In other words, about 30 percent of the between-district variance in mean achievement is accounted for by district home environment/motivation. Mean achievement, however, still varied significantly across districts (χ^2 = 13013, df=494, p<.001). This indicates that more terms need to be added to the between-district model in order to account for additional variation in district mean achievement (i.e., the explanatory model is still misspecified).

Table 3 displays the results for an HLM model using the variables from the OLS district-level analysis with θ coefficient values, standard errors and t-ratios:

Table 3
HLM Results for District Mean Achievement

Residual Parameter Variance 4.33
R Square .62

VARIABLE	θ	S.E.	T	p
DACH86	.492	.039	12.730	.000
AVGHOMTV	1.726	.3?6	4.361	.000
AFDCPCT	-.085	.017	-4.995	.000
LNADM	-.436	.173	-2.516	.012
CLSIZ	-.361	.111	-3.266	.001

In this model after accounting for the effects of DACH86, AFDCPCT, AVGHOMTV, LNADM and CLSIZ, the residual parameter variance is reduced to 4.33, a reduction of 62 percent over the unconditional model with no district effects included. The θ effect for DACH86 indicates a highly significant association with district mean achievement. The other four predictors, to a lesser extent, are also significantly rel... d to explaining variation in district mean achievement as indicated by their θ coefficients. In this model, however, residual parameter variance in mean achievement still varied significantly across districts (χ^2 = 3352.8, df=490, p<.001).

The strong effect of DACH86 on district mean achievement mitigates the effects of the other district-level predictors.[5] In an attempt to assess the potential effects of these variables, a model was formulated without DACH86. The results of this model are shown in Table 4.

Table 4
HLM Results for Model without Prior Ability

Residual Parameter Variance	6.31
R Square	.44

VARIABLE	θ	S.E.	T	p
AVGHOMTV	4.051	.406	9.983	.000
AFDCPCT	-.145	.019	-7.643	.000
LNADM	-.573	.200	-2.865	.005
CLSIZ	-.530	.127	-4.170	.000

In this new model, the θ coefficients for the four district-level variables increase considerably. The estimated residual parameter variance, τ, for this model is 6.31 which represents a reduction of 44 percent over the unconditional model.[6] Mean achievement, as in the previous case, still varied significantly across districts (χ^2 = 6302.7, df=491, p<.0001).

The remaining set of HLM analyses focused on modeling the variability in within-district slopes as a function of district-level characteristics. In this manner, the distribution of academic achievement was studied both within and across districts. The modeling strategy employed

[5]For proper model specification, however, prior ability needs to be included in order to control for the nonrandom manner in which students are placed into schools and districts. Otherwise, the other variables take on importance which may merely be attributable to initial student differences.

[6]In contast, a between-district model containing only DACH86 resulted in a reduction of 54 percent in terms of explaining parameter variance.

here built upon the previous analysis of district mean achievement. An unconditional model was proposed whereby distric* mean achievement (BASE) was modeled along with the slopes for ACH86 and HOMEMOTV:

$Y_{ij} = \beta_{j0} + \beta_{j1}(ACH86) + \beta_{j2}(HOMEMOTV) + \beta_{j3}(CEPRM88)$

In this model where β_{j0} = BASE achievement, ACH86, HOMEMOTV and CEPRM88 were centered around their respective district means and represent the "differentiating effect" of each variable within district *j*. In the unconditional model, the questions of interest centered around determining, on the one hand, whether there was a significant regression effect of the *k* student-level variables on academic achievement within districts, as well as the extent of variation of these effects across districts. In the unconditional between-district model, each OLS regression coefficient, β_{jk} was in turn modeled as:

$$\beta_{jk} = \mu_k + U_{jk} \text{ for } k = 0,1,2,3. \tag{10}$$

The μ_k represent the fixed main effects (constant for each district) while the U_{jk} are the random effects which vary from district to district. These random effects represent the unique increment to the slope contributed by district *j*. The results for the unconditional model are displayed in Table 5.

The θ coefficients provide estimates of the mean fixed effects. Each of these mean fixed effects is statistically significant at the .05 level, indicated by the individual t-statistics testing the hypothesis (e.g., H_0: $\theta_k = 0$) of whether the average within-district coefficient = 0.

The results for the random effects U_{jk} indicate heterogeneity of regression for the four coefficients across districts. The BASE, ACH86, HOMEMOTV AND CEPRM88 effects all varied significantly across districts as indicated by the results of the Chi-square tests.[7] For example, in the case of the ACH86 slope, a test of the null hypothesis H_0: $\tau_{ACH86} = 0$ yields a test statistic

[7]Note that these Chi-square tests are conceptually equivalent to testing for homogeneity of regression in an ANCOVA model. The distinction here is that HLM permits an explanatory model to be posited which may account for the random variability in slopes across districts.

of 1903.1 which is then compared to a Chi-square critical value with 491 degrees of freedom. The result for HOMEMOTV (p=.048), on the other hand, suggests that the residual parameter variance for this effect is quite close to 0.

Table 5
HLM Results for Unconditional Model

Fixed Effects	**θ**	**S.E.**	**T**	**p**
BASE, β_0	82.8529	.1613	513.732	.000
ACH86, β_1	.7087	.0067	105.572	.000
HOMEMOTV, β_2	.7962	.0244	32.642	.000
CEPRM88, β_3	-3.7233	.1347	-27.649	.000

Random Effects	**Estimated Parameter Variance**	**df**	**χ^2**	**p**
BASE ACHIEVEMENT	12.2228	491	47763.0	.000
ACH86 SLOPE	.0156	491	1903.1	.000
HOMEMOTV SLOPE	.0093	491	544.4	.048
CEPRM88 SLOPE	5.2584	491	1464.3	.000

RELIABILITIES OF DISTRICT-LEVEL RANDOM EFFECTS	
BASE ACHIEVEMENT	= .946
ACH86 SLOPE	= .639
HOMEMOTV SLOPE	= .019
CEPRM88 SLOPE	= .493

Information about the reliability of the random effects in the model is also provided. These reliability indicators are derived from the ratio of estimated parameter variance in each regression coefficient, $\hat{\sigma}^2(\beta_{jk})$ to the total observed variance in the estimated OLS slopes, $\hat{\sigma}^2(\beta_{jk})$ + $\hat{\sigma}^2(\hat{\beta}_{jk} | \beta_{jk})$. The estimate for BASE is highly reliable, .946. This indicates that almost all of the total observed variance in base achievement is potentially explainable by district-level factors. The regression coefficients are less reliable, ranging from a high of .639 for ACH86 to a low of .019 for HOMEMOTV. In this latter instance, approximately 98 percent of the observed variation in HOMEMOTV is attributable to sampling variance and not explainable by district-level characteristics.[8]

The final step in the multilevel analysis involved fitting a model to demonstrate the capability of the HLM program to model the effects of district-level covariate and policy manipulable variables on within-district slopes. In this final model, a "sensitivity" analysis was conducted to determine the most economical set of covariates to accompany the policy variables of LNADM, EXPEND and CLSIZ. In the interests of parsimony and interpretability, only these three policy variables were modeled.[9] In this model, the estimated residual parameter variance for HOMEMOTV was close to 0 indicating that the homogeneity hypothesis of residual variance for this parameter could be sustained and that any remaining variance could be attributed to sampling variability. It was thus decided to "fix" the residual variance in the HOMEMOTV slope to 0, whereby HOMEMOTV was treated as a fixed component with only an intercept term and no residual variation to explain. The results from the final explanatory model are presented in Table 6.

In terms of district mean achievement, both LNADM and CLSIZ had negative θ coefficients, replicating the results from the ordinary least squares district-level analysis. EXPEND had a negligible negative effect on district achievement. In terms of the ACH86 slope, CLSIZ had a small positive effect on the differentiating effect of prior ability on achievement. Within the HLM formulation of centering within-group variables around their group means, the intercept β_{j0} now represents group mean achievement. This choice of metric also allows for unambiguous statements to be made concerning the within-group coefficients

[8]A potential explanation for the low reliability of the home environment-motivation effect is the lack of variation in HOMEMOTV between and within districts. Another possible explanation is collinearity among the within-district slopes. In a model without ACH86, the parameter variance for HOMEMOT increased from .0093 to .5923 and the reliability indicator for the slope similarly rose from .019 to .428.

[9]TCHLEVEL was also initially considered, but this variable was subsequently dropped from the analysis due to its possible suppressor effect (see results of OLS district-level analysis).

Table 6
HLM Results for Explanatory Model of District-Level Effects

Fixed Effects	θ	S.E.	T	p
FOR BASE, β_0				
BASE	31.4679	2.9567	10.643	.000
DACH86	.5737	.0322	17.835	.000
AVGHOMTV	1.8172	.4356	4.172	.000
AFDCPCT	-.0498	.0146	-3.415	.001
LNADM	-.4658	.1760	-2.647	.009
EXPEND	-.0002	.0002	-.889	.374
CLSIZ	-.3666	.1166	-3.144	.002
FOR ACH86, β_1				
BASE	.3845	.1306	2.945	.004
AVGHOMTV	.0546	.0207	2.633	.009
PCTNW	-.0012	.0005	-2.190	.028
DCEPRM88	-.0044	.0006	-7.147	.000
DCUTTELS	.1929	.0383	5.042	.000
LNADM	-.0100	.0093	-1.078	.281
EXPEND	.00001	.00001	.785	.433
CLSIZ	.0122	.0060	2.043	.041
FOR HOMEMOTV, β_2*				
BASE	.8147	.0232	35.057	.000
FOR CEPRM88, β_3				
BASE	-5.0274	1.8118	-2.775	.006
LNADM	-.0928	.1969	-.471	.637
EXPEND	.0006	.0002	2.916	.004
CLSIZ	.0141	.1377	.102	.919

* - The residual variance for this parameter has been set to zero.

Random Effects	Estimated Parameter Variance	df	χ^2	p
BASE ACHIEVEMENT	5.5329	485	8138.0	.000
ACH86 SLOPE	.0108	484	1773.0	.000
CEPRM88 SLOPE	5.0019	488	1480.4	.000

and their relationship to the group mean on the outcome measure (Bryk et al., 1988). Consequently, the differentiating effect of a within-district slope coefficient can be interpreted here as the degree to which differences in a within-district (e.g., prior ability) relate to differences in fifth-grade achievement. Larger class size, thus, tended to magnify the gap between students scoring low and high on third grade achievement. Conversely, smaller class size resulted in reducing this gap by preventing the low achievers from falling further behind in terms of achievement. Figure 1 depicts this relationship between the district-level effect of class size and the ACH86 slope. The ACH86 slope modeled with CLSIZ is steeper than the unconditional ACH86 slope indicating that larger class size has an effect of increasing the gap between low and high students on third grade achievement.

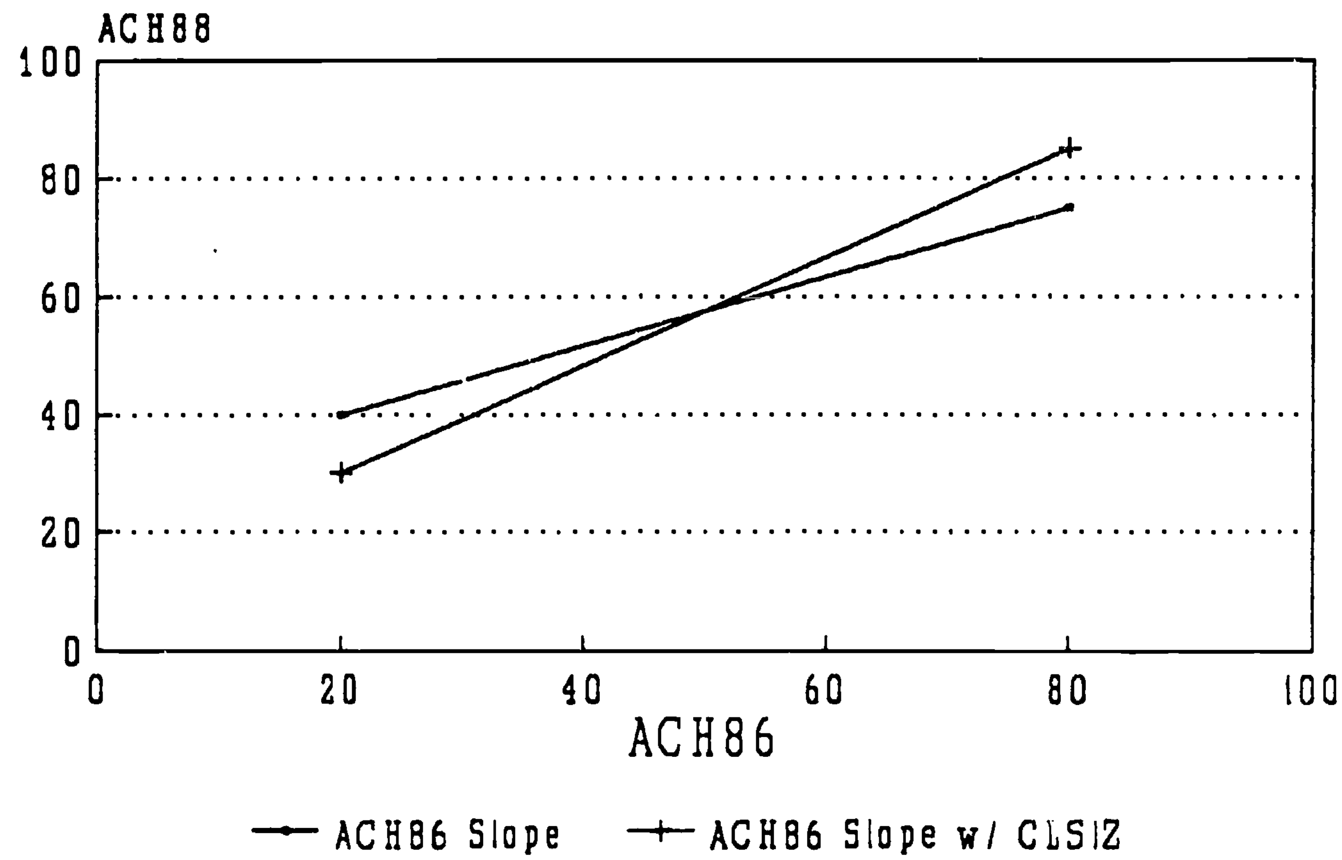

Figure 1
Effect of Class Size on ACH86 Slope

LNADM and EXPEND, on the other hand, exhibited no significant effects on either decreasing or increasing the differentiating effect of ACH86 on achievement. In addition, the policy variables LNADM and CLSIZ were negligibly related to explaining the variation in the CEPRM88 slope. EXPEND, however, did show a small positive effect in reducing the differentiating effect of compensatory education performance on achievement. .he higher a district's level of instructional expenditures per pupil, the less of a gap between those students receiving and not receiving remedial services in reading and math. This effect (θ = .0006) may be too minute, however, to have any impact in a practical significance sense.

The explanatory model accounted for 54.7% of the parameter variance in BASE, 30.8% in ACH86 and 4.9% in CEPRM88. The Chi-square statistics for the homogeneity of regression test for the three random terms in the model was significant indicating that this model was inadequate in terms of explaining the parameter variance among districts.[10]

Summary

Policy relevant variables, thus, did not make a major contribution in terms of explaining within-district variability. Proper model specification in this case entails finding additional or better measures of district-level indicators which could explain the differences in base achievement and slope variation across school districts. Use of a multilevel modeling procedure, such as HLM, allows one, however, to draw a broader range of policy inferences than one would under a conventional regression framework. That is, one can identify those factors which explain not only variation in the outcome measure, but within-group structural relationships as well, such as within-district slopes. For example, (see Table 6), the effects of district size and expenditures can take on different interpretations. In terms of district mean achievement, district size had a moderate negative effect while the effect of expenditures was negligible. However, district size had no effect on increasing the gap between students either scoring low and high on third grade achievement (ACH86 slope) or those receiving and not receiving remedial services in reading and math (CEPRM88 slope). Thus, from one set of results, one could make a case against school district consolidation with the argument that district size has a negative effect on district achievement. By examining relationships within

[10]These Chi-square statistics are only indications of the statistical fit of a model. Due to the presence of a few large school districts in this study producing very small standard errors, these statistics can become unduly inflated. Consequently, it may be preferable to examine the "substantive" fit of a model rather than rely strictly on statistical criteria.

districts, however, one could also argue for consolidation by showing that district size has a negligible effect on low achievers or those students receiving compensatory instruction falling further behind in achievement. To make informed policy decisions requires, therefore, as much data-based information as possible which can be brought to bear on an issue.

Similarly, additional information is available for policy purposes from an HLM analysis with regard to the class size variable (CLSIZ). As shown in Table 6, class size has a negative effect on district mean achievement. The HLM analysis reveals, in addition, a small effect of class size in increasing the differentiating effect of prior ability (ACH86 slope) on achievement. This supplemental information could be used to inform the policy debate on decreasing class size.

How does an HLM analysis employing multilevel data differ from conventional analyses conducted at the individual or district level? The results from Tables 1 and 3 show a striking similarity between the estimates obtained from the ordinary least squares regression district-level analysis and from the HLM analysis of district mean achievement. This is not surprising given the fact that in both instances mean district achievement is being predicted by the same set of variables. The multilevel HLM analysis, however, accounted for a greater share of the explained variance in achievement (62 percent) than the OLS analysis at the district level (51 percent). This is explained by the fact that the hierarchical linear modeling procedure partitions the variance in district means into parameter variance and sampling variance as opposed to conventional regression analysis which does not make this distinction. Thus, in the HLM case, the set of five predictor variables explained 62 percent of the "true" differences in district achievement potentially explainable, which were not attributable to sampling variability.

This particular distinction between a multilevel modeling procedure such as HLM and ordinary least squares estimation procedure is critically important in terms of accurately measuring the extent of school or district effectiveness. Failure to properly partition the variation between groups into that portion capable of explanation as opposed to noise has led to serious underestimation of school effects in the past (see Willms, 1984).

The above results suggest various implications for policy research, particularly in terms of how policy makers can potentially be misled by the results of inappropriate methodological models given the hierarchical nature of educational data. For one, the importance of correctly partitioning variability among schools or school districts has clear implications for educational policy research. As indicated in the results only a small portion of variability in student achievement is potentially explainable by group-level factors. Analyses employing district-level aggregate data, thus, are only capable of explaining a finite portion of the variation in achievement. District-level analyses explain **district-level** achievement.[11] The district-level model in this study with an R^2 of .51 is only explaining, in fact, 51 percent of the potentially explainable portion (in this case six percent) of the variation in student achievement. Adequate model specification, thus, depends on more than the inclusion of all relevant predictor variables related to the outcome measure. Failure to take into account the fact that students are grouped within schools and/or school districts, for example, can greatly obscure the findings of an analysis solely employing aggregate-level data.

Can policy questions be informed by results of analyses conducted with individual-level data? The individual level OLS model, in contrast to the district-level OLS analysis, had a higher R^2 of .64 primarily due to the stronger relationship between third and fifth grade achievement at the student level. Individual level models, however, while explaining variation in **student-level** achievement, are incapable of modeling district-level factors (policy variables and contextual effects) without biasing other effects in the model. Simultaneously modeling variables measured at different levels of aggregation results in inadequate estimation of model

[11]Of that portion, it is crucial to identify those manipulable policy variables which have a potential impact on individual student achievement (Langbein, 1977).

parameters due to the pattern of intercorrelations between the individual and district-level effects (see Aitkin & Longford, 1986; Boyd & Iversen, 1979, for further discussion of this issue).

The purpose of the above comparisons was not to demonstrate the superiority of one level of analysis over the other, but rather to show that comparisons of analyses conducted at different levels of aggregation may not be meaningful. The question of which level of analysis to choose is clearly the wrong question to ask. The emphasis, instead, should be on developing a model to fit how the data is generated.

> The focus of an investigation of educational effects should be on the proper specification of the substantive analytical model(s) rather than on making a choice among competing units of analysis (Burstein, 1980, p. 161).

Moreover, variables measured at the individual student level often represent different constructs from their school- or district-level counterparts (Cronbach, 1976). For example, the student-level indicator of SES in these data, HOMEMOTV, measures the propensity of an individual student to do well on achievement tests. The district-level aggregate, AVGHOMTV, even though it may have the same functional relationship with district-level achievement as HOMEMOTV does with individual-level achievement, takes on a different meaning in terms of its relationship with other district-level indicators such as per-pupil spending. The average district SES score may reflect more the fact that higher SES families enroll their children in better schools. It does not guarantee, however, that a student located in a high SES district will do well on achievement. Nor can a conclusion be drawn from an individual's high SES as to how that student's district will perform in terms of achievement. The relevant question to ask in this case is the effect of average SES composition of a school district on within-district variation in individual SES.

These analyses have demonstrated, moreover, the importance of specifying prior ability in a model to reduce the bias due to student self-selection.[12] Removing prior ability from the

[12]Conditioning on prior ability reduces but cannot eliminate initial differences in academic achievement.

district-level model reduced the R^2 by 30 percent and affected the magnitude of the other parameters in the model. In the individual-level model, removing prior ability reduced the R^2 by 36 percent and caused the remaining effects to increase dramatically in size.

Furthermore, analyses employing proxies such as SES for incoming student ability are simply inadequate for addressing this bias. In the HLM model of district achievement, prior ability (DACH86) accounted for 80 percent more explanatory power than SES (AVGHOMTV) in predicting district mean achievement. The bias introduced into a predictive model when controlling solely for student SES and ignoring prior ability has been extensively documented in the literature (see Gray, 1989, for examples).

As a recommendation for state data collection efforts, stronger and more reliable measures of SES are needed, both at the individual and district level. Indicators at the individual level based on student self-report data are subject to problems of unreliability endemic to questionnaire data. The student self-report data collected in this study are particularly prone to this problem considering the age level of the students. The extremely low reliability estimate for the HOMEMOTV slope in the HLM analysis (see Table 5) underscores this point. At the district level, a factor-weighted composite index would provide an improvement over the use of three disparate indicators for SES. States interested in assessing district effectiveness would be well-served by systematically collecting this information.

In short, the above discussion serves to emphasize the importance of proper model specification, both from the perspective of adequately controlling for student self-selection, and from capturing the educational processes underlying the data through the delineation of a multilevel analytical model. The statistical model proposed, however, should match as closely as possible the substantive model responsible for generating the data within a hierarchical context. In this manner, a multilevel analysis guided by estimation techniques employed in this study allows the educational researcher to not only avoid the problems of aggregation bias and model misspecification, but to identify the factors responsible for explaining the variation in individual academic achievement.

The finding in this study of only six percent of the variance in achievement lying between districts replicates a similar finding from the Coleman et al., (1966) study, whereby the majority of variance was found within rather than between schools. Similarly, studies conducted by Gray (1989) found very small proportions of the variance in student achievement existing among British local education authorities. In this present study the lack of a strong grouping effect at the school district level would lead one instead to look at potential school-level factors to explain the variability in student achievement. Better and more complete estimates of student attributes to redress the problems of missing data and non-response, as well as the employment of longitudinal data to measure long-term educational effects, are other concrete examples of developing more adequately specified models of student achievement. The emphasis, in any case, should be on the development of simple, parsimonious models for ease of interpretation.

It is not expected that this approach using hierarchical linear modeling can provide a conclusive conceptual answer to the question of correct model specification of student achievement. Due to the non-random manner in which students are grouped into schools and districts, model misspecification in terms of biased parameter estimates is likely to continue to be a problem, no matter how sophisticated the statistical methodology employed. Model misspecification and biased estimation can be reduced, however, with the development of stronger conceptual models. The sophistication of our conceptual modeling efforts has unfortunately lagged behind recent developments in statistical methodology. By specifying the process through which variables measured at different levels are related, however, multilevel analysis offers hope of an improvement over conventional regression analyses confounded by aggregation bias and model misspecification.

REFERENCES

Aitkin, A., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society,* (Series A), *149,* 1-43.

Bernstein, L.S. (1990). *The application of hierarchical linear modeling to multilevel student achievement data.* Unpublished doctoral dissertation, University of Pittsburgh.

Bidwell, C.E. & Kasarda, J.D. (1975). School district organization and student achievement. *American Sociological Review, 40,* 55-70.

Blalock, H.M. (1964). *Causal inferences in nonexperimental research.* Chapel Hill: University of North Carolina Press.

Boyd, L.H. & Iversen, G.R. (1979). *Contextual analysis: Concepts and statistical techniques.* Belmont, CA: Wadsworth.

Bryk, A.S., Raudenbush, S.W., Seltzer, M., & Congdon, R.T. (1988). *An introduction to HLM: Computer program and users guide.* University of Chicago, Department of Education.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education, 8,* 158-233.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D. & York, R.L. (1966). *Equality of educational opportunity.* Washington, D.C.: Office of Education, U.S. Department of Health, Education and Welfare.

Cooley, W.W. (1978). Explanatory observational studies. *Educational Researcher, 7*(9), 9-15.

Cooley, W.W., Bond, L., & Mao, B-J. (1981). Analyzing multilevel data. In R.A. Berk (Ed.), *Educational evaluation methodology: The state of the art.* (pp. 64-83). Baltimore: The Johns Hopkins University Press.

Cronbach, L.J. (1976). *Research on classrooms and schools: Formulation of questions, design and analysis.* Occasional paper of the Stanford Evaluation Consortium.

Friedkin, N.E. & Necochea, J. (1988). School system size and performance: A contingency perspective. *Educational Evaluation and Policy Analysis, 10,* 237-249.

Geske, T.G. (1983). Education finance policy: A search for complementarities. *Educational Evaluation and Policy Analysis, 5,* 83-96.

Glasman, N.S. & Biniaminov, I. (1981). Input-output analyses of schools. *Review of Educational Research, 51,* 509-539.

Gray, J. (1989). Multilevel models: Issues and problems emerging from their recent application in British studies of school effectiveness. In R.D. Bock (Ed.), *Multilevel analysis of educational data.* (pp. 127-145). San Diego: Academic Press.

Guthrie, J.W. (1979). Organizational scale and school success. *Educational Evaluation and Policy Analysis, 1,* 17-27.

Hannan, M.T. (1971). Problems of aggregation. In H.M. Blalock (Ed.), *Causal models in the social sciences.* (pp. 473-508). Chicago: Aldine-Atherton, Inc.

Langbein, L.I. (1977). Schools or students: aggregation problems in the study of student achievement. In M. Guttentag (Ed.), *Evaluation Studies Review Annual, Vol. 2.* (pp. 270-298). Beverly Hills, CA: Sage.

Longford, N.T. (1989). Appendix 1. Variance component analysis. In D.J. Smith & S. Tomlinson (Eds.), *The school effect.* (pp. 309-319). London: Policy Studies Institute.

Madaus, G.F., Airasian, P.W., & Kellaghan, T. (1980). *School effectiveness: A reassessment of the evidence.* New York: McGraw-Hill.

Mosteller, F. & Moynihan, D.P. (1972). *On equality of educational opportunity.* New York: Random House.

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

Raudenbush, S.W. & Bryk, A.S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. In E.Z. Rothkopf (Ed.), *Review of Research in Education, Vol. 15.* (pp. 423-475). Washington, D.C.: AERA.

Summers, A.A. & Wolfe, B.L. (1977). Do schools make a difference? *The American Economic Review, 67,* 639-652.

Turner, R., Camilli, G., Kroc, R., & Hoover, J. (1986). Policy strategies, teacher salary incentive, and student achievement: An explanatory model. *Educational Researcher, 15*(3), 5-11.

Walberg, H.J. & Fowler, W.J. (1987). Expenditure and size efficiencies of public school districts. *Educational Researcher, 16,* 5-13.

Willms, J.D. (1984). School effectiveness within the public and private sectors. *Evaluation Review, 8,* 113-135.