REGRESSION BASED ANALYSES OF MULTILEVEL EDUCATION DATA

Leigh Burstein
Michael David Miller

CSE Report No. 175

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

This paper focuses on three topics within the domain of regression-based analyses of multilevel data from quasi-experiments and field studies in educational research and evaluation. The paper begins with a discussion of the general question of choice of unit of analysis or what may be the more appropriate question of choice of analytical model. After discussing this issue and providing empirical illustrations of the importance of knowing the question of interest, two additional topics will be considered: the use of within-group slopes as indices in between-group analyses and the estimation of within-group dependency and its role in analyses of multilevel data. These latter topics reflect important substantive concerns in school-based non-experimental investigations.

Overall, we believe that the major technical complication in the analysis of multilevel data from quasi-experiments and field studies is the inability of educational researchers to develop adequate theories about educational processes within groups (classrooms and schools) and to develop adequate methodology for analyzing the educational effects of such processes. The material presented here reflects an attempt to systematize the investigation of two important indices of within-group processes.

Choice of Units of Analysis and/or Choice of Analytical Model

Efforts to identify the effects of education (e.g., Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld and York, 1966) on pupil performance have suffered from the complications caused by the multilevel character of educational data. Schools are aggregates of their teachers, classrooms and pupils, and classrooms are aggregates of the persons and processes within

them. This being the case, the effects of education can exist both between and within the units at each level of the educational system. Yet the majority of studies of educational effects have restricted attention to either overall between-student, between-class, or between-school analyses.

Cronbach (1976) argued that the majority of studies of educational effects carried out thus far conceal more than they reveal, and that "the established methods have generated false conclusions in many studies" (p. 1). His concern is foreshadowed in the educational literature by the exchange among Wiley, Bloom, and Glaser as recorded in Wittrock and Wiley (1970), and by Haney's (1974) review of the units of analysis problems encountered in the evaluation of Project Follow Through.

Research on the differences between multiple regression models at different levels of aggregation (Burstein, 1975 , 1978; Hannan and Burstein, 1974; Hannan and Young, 1976a; Feige and Watts, 1972)and on the analyses of school effects at different levels (Burstein, Fischer, and Miller, 1978; Burstein and Smith, 1977; Comber and Keeves, 1973; Hannan, Freeman, and Meyer, 1976; Keesling and Wiley, 1974) indicates that (a) there are substantial differences in the magnitudes of regression coefficients across levels for specific models; (b) different variables enter the models at different levels; and (c) aggregation generally inflates the estimated effects of pupil background and decreases the likelihood of identifying teacher and classroom characteristics that are effective. The results cited above are not very comforting for the researcher who wishes to draw conclusions about educational processes at one level but is constrained to analysis at a different level.

When faced with the analysis of multilevel data, most researchers have tried to make a choise among alternative units of analysis on the basis of theory or statistical considerations. Unfortunately, those who resort to theory either reject plausible alternative models (Brophy, 1975; Bloom, 1970; Stebbins, St. Pierre, Proper, Anderson and Cerva, 1977; Wiley, 1970) or find themselves unable to choose (Cline, Ames, Anderson, Bale, Ferb, Joshi, Kane, Larson, Park, Proper, Stebbins, Stern, 1974; Haney, 1974). Picking the appropriate unit on the basis of statistical considerations can also leave the choice unresolved due to competing alternatives (Burstein and Smith, 1977; Glendening, 1976; Haney, 1974).

Haney (1974) has elaborated the range of alternative considerations in the contextof the evaluation of Project Follow Through. He cites four general types: the purpose of the evaluation (questions to be addressed), the evaluation design (nature of treatments, independence of units and treatment effects, appropriate size), statistical considerations (reliability of measures, degrees of freedom, analysis techniques), and practical considerations (missing data, policy research, multiple year comparisons, economy). Haney was unable to choose among untis because the purpose of the evaluation dictated the child as the unit but the unit of treatment was the classroom; moreover, the multiyear character of Follow Through made classrooms impractical as units of analysis. And, since there was no random assignment at any level and the comparison children were not equivalent to treatment children, these considerations offered no relief.

Apparently, thinking of multilevel analyses simply as problems in the choice of a unit of analysis is inadequate. Phenomena of importance occur at all levels and need to be described and subjected to inference-making. (Burstein and Linn, 1976; Cronbach, 1976). Once again, Haney's arguments are succinct and to the point:

> Investigators ought to have a strong bias for studying various properties of the educational system at the level at which they occur; . . .variation in attributes of interest ought to be studied at those levels (or between those units) at which it does (or is expected to) occur. . . If the hypotheses are explicitly stated in terms of mathematical models, the impact of shifting levels of analysis from one unit of analysis to another will be much more easily assessed than if they are not (1974, pp. 96-97).

These arguments cited by Haney serve as justification for the research we describe throughout this paper.

Decomposition into Between-Group and Within-Group Effects

A variety of competing points can be cited as traditional justification for the choice of either pupils or groups (classrooms, schools, etc.) as the appropriate unit of analysis in studies of educational effects. Generally arguments cited are compelling and virtually irreconcilable if a choice of either pupil or group as the only unit is required. The multilevel character of educational data warrants analytical strategies tailored to the identification of educational effects at and within each level of the educational system. Moreover, the complexity of the choice depends on the type of study being conducted as well as the types of outcomes and processes under investigation.

Even in the simplest case, once the existence of specific group membership is acknowledged (e.g., instruction from a specific teacher), any measure that varies over pupils can be decomposed into its between-group and within-group components. For example, if we consider the posttest

or outcome performance, Y_{ij}, of pupil j in class i (j = 1,..., n persons per class; i = 1,..., k classes; for simplicity we assume equal-size classes) and the performance level X_{ij}, of the pupil prior to entering the class (i.e., the pretest or some measure of entering ability), then the relation of X_{ij} to Y_{ij} can be decomposed into between-class and within-class components (Burstein, Linn, and Capell, 1978; Cronbach, 1976):

$$
\begin{aligned}
Y_{ij} - \bar{Y}_{..} &= \beta_b(\bar{X}_{i.} - \bar{X}_{..}) && \text{Predicted Between-Class} \\
&+ \bar{Y}_{i.} - \beta_b(\bar{X}_{i.} - \bar{X}_{..}) && \text{Adjusted Between-Class} \\
&+ \beta_w(X_{ij} - \bar{X}_{i.}) && \text{Pooled Within-Class Slope} \\
&+ (\beta_i - \beta_w)(X_{ij} - \bar{X}_{i.}) && \text{Specific Within-Class} \\
&+ \varepsilon_{ij} && \text{Specific Residual Associated with Person ij}
\end{aligned}
$$

In the above equation, β_b is the between-class slope from the regression of $\bar{Y}_{i.}$ on $\bar{X}_{i.}$, β_w is the pooled within-class slope from the regression of $(Y_{ij} - \bar{Y}_{i.})$ on $(X_{ij} - \bar{X}_{i.})$ <u>across all</u> classrooms, and the β_i are the <u>specific</u> within-class slopes from the regression of Y_{ij} on X_{ij} within the i classrooms.

The possible substantive interpretations of specific components and sets of components are important here. (See particularly descriptions of alternative analytical models and the section on slopes as indices). The key elements are the between-class slope, the adjusted between-class effect, the pooled within-class slope, and the specific within-class slopes.

Often, Equation (1) can be modified so that we have a global measure, T_i, of classes (e.g., class membership, teacher quality, or treatment-group) rather than the aggregation of individual scores represented by $\bar{X}_{i.}$.

In what follows we shall refer to the effects associated with either $\bar{X}_{i.}$ or T_i as class effects without loss of generality.

One useful treatment of the multilevel analysis was provided by Cronbach (1976). A succinct statement of Cronbach's justification for his proposed analysis is that the usual overall between-student analysis combines two kinds of relationships--those operating between collectives (reflected in β_b and adjusted class effects) and those operating among persons within collectives (reflected in β_w and β_i)--into a composite that is rarely of substantive interest (Cronbach, 1976, pp. 10.3ff.). Cronbach reminds us that β_t, the overall between-student coefficient from the regression of Y_{ij} on X_{ij},

$$(2) \quad Y_{ij} - \bar{Y}_{..} = \beta_t (X_{ij} - \bar{X}_{..}) \quad ,$$

has been shown by Duncan, Cuzzort, and Duncan (1961, p. 66) to be a composite of β_b and β_w:

$$(3) \quad \beta_t = \eta_X^2 \beta_b + (1 - \eta_X^2)\beta_w$$

where η_X^2 is the intraclass correlation or correlation ratio of X. Cronbach (1976; Cronbach and Webb, 1975) goes on to recommend that between-group effects and individuals-within-group effects should be examined separately.

In its most parsimonious form, Cronbach would examine the following:

$$(4) \quad \underline{\text{Between Groups}}: \quad \bar{Y}_{i.} - \bar{Y}_{..} = \beta_{\bar{y}t} T_i + \beta_b (\bar{X}_{i.} - \bar{X}_{..})$$

where the $b_{\bar{Y}T}$ is the effect of teachers on mean outcomes after controlling for between-class differences in inputs.

$$(5) \quad \underline{\text{Pooled Within-Groups}}: \quad Y_{ij} - \bar{Y}_{i.} = \beta_w (X_{ij} - \bar{X}_{i.})$$

Thus, Cronbach's primary concerns are with the adjusted collective effects of instruction as reflected by the adjusted class mean outcomes and with the overall redistributive properties of classroom instruction as reflected by the pooled within-class regression, β_w.

Empirical Results from Multilevels of IEA Data

For the time being, we focus on the two estimators of most interest to Cronbach, between-group regression coefficients and the corresponding pooled within-group coefficients. Recent empirical analyses of data from the IEA Six Subject Survey (Burstein, Fischer, and Miller, 1978) dramatically demonstrate the distinct differences in interpretation when one moves from a between-school to a within-school analysis. This study investigated the factors influencing educational achievement in twenty-one countries, considering six subject areas (Science, Reading Comprehension, Literature, Civics Education, English as a Foreign Language, and French as a Foreign Language) at three age levels (basically, 10 year-olds, 14 year-olds, and students in their preteritary year). Over 700 student, teacher, and school characteristics were measured.

In an investigation of educational effects models for 14 year-olds from the U.S. and Sweden in the IEA science achievement study (Table 1) we found that the effects of family background on science achievement were substantial, as usual, in the between-schools analysis of U.S. data but much smaller in Sweden. In fact, for 14 year-olds, R^2_{Total} was larger than $R^2_{\text{Between-schools}}$ in Sweden, which would be atypical for analysis of U.S. data.

In contrast, the effects of family background in the pooled-within-school analyses for the U.S. were substantially smaller and were essentially the same as the effects found in the within-school analysis

for Sweden. One possible substantive explanation for these findings is that the two types of analyses reflect, on one hand, distinctions between the countries in the political order governing the distribution of pupil backgrounds and school resources (i.e., the predominance of local control and community determination of school resources in the U.S. vs. national control and a policy of uniformity of resources for Sweden) and, on the other, similarity between countries in the operation of the social order within schools (i.e., interpersonal allocations of rewards within an institution).

There are further substantive quesions that the above example might address, but the methodological point is clear: different types of analysis of multilevel data address different questions and typically research on schooling asks questions at multiple levels.

Within-Group Slopes as Indices in Between-Group Analyses

Once it is determined that the questions of interest and/or statistical considerations warrant analyses of aggregated data, the types of between-group effects one expects to find remain to be specified. In particular, when one's purpose is to determine factors affecting pupil performance, it is possible that analyses of between-group (class, school, etc.) means can hide important differences in the within-group distribution of pupil outcomes and educational inputs.

Several aspects of current scho-ling practices lead us to expect that within-school and within-class distributions of pupil performance vary. First, schools (classes) do differ in the distribution of educational

performance. Moreover, schools with the same mean outcome often exhibit different distributions of performance within school. An analysis of means alone could not be expected to account for such distributional differences.

Second, a variety of educational theories about the effects of specific schooling practices on within-group behavior argue for an examination of distributional properties other than group means. Obviously, at least the variability of performance is of interest in studies comparing individualized, competency-based, or open educational instructional programs with more traditional instructional practices. Also, research on the interaction between teaching style and learning style would lead one to expect variability of outcomes for pupils with similar entering characteristics and preferences taught by teachers with differing instructional styles.

Finally, the idea of using distributional characteristics in addition to the mean as criterion measures has been shown previously to merit consideration (Lohnes, 1972; Klitgaard, 1975; Brown and Saks, 1975). Lohnes (1972) found that standard deviations and skewness indices added to the explanatory power of means in his analyses of data from the Cooperative Reading Project. Klitgaard (1975) and Brown and Saks (1975) found that school and school district standard deviations exhibited more significant relations with school characteristics than did school and school district means.

Though they sought answers to different questions and used different methodologies, Lohnes, Brown and Saks, and Klitgaard apparently share our belief that educational outcomes are multifaceted and incompletely measured by single group averages. There also seems to be consensus that educational theory can be developed which will link pupil entering characteristics and

and characteristics of the educative process to distributional properties of educational outcomes.

We (Burstein and Linn, 1976; Burstein, Linn and Capell,, 1978) have elaborated a theory for the use of within-group slopes of outcomes on inputs as a criterion in educational effects studies. Wiley (1970) may have been the first to suggest this strategy.

Our justification for considering within-group slopes as outcomes derives much of its impetus from research on aptitude-treatment interactions (Snow, 1976) and from evidence of slope differences among colleges (Rock, Baird, and Linn, 1970). In its simplest form, we expect that different combinations of teachers and instructional practices will result in varying distributions of educational outcomes for pupils with similar entering characteristics. For example, it might be hypothesized that there are teachers who are equally effective in obtaining mean performance, but yield varying slopes because some teachers use compensatory instructional practices which emphasize the improved performance of lower-ability students while others allow each child to learn at his/her own rate. (We would expect a flatter slope in the former case than in the latter.)

Burstein and Linn (1976; Burstein, Linn and Capell, 1978) compared alternative analytical models for identifying educational effects for sets of hypothetical classrooms with heterogeneous slopes. The key findings were that, for the conditions studied, heterogeneous within-class slopes were shown to make important differences in identified effects, ones which were not swamped by sampling variability in the estimation of slopes, and certain analytical strategies exhibited good properties even in the presence of heterogeneity.

Although within-group slopes are conceptually appealing indices of educational effects, three points warrant further examination. First, it must be determined that slopes are sufficiently stable. Second, it must be demonstrated that slopes are potentially distinct from other group indices (e.g., pre and posttest means and standard deviations) in realistic situations. Finally, there have to be realistic cases in which slopes are related to school and class characteristics after controlling for other background measures and other indices of group outcomes. We have already begun to investigate these points (see below).

Stability of Slopes

The sampling variability of within-group slopes is substantially greater than that of the mean. For small samples, e.g., the size of a classroom, the sampling error of a slope is so large that it is questionable whether real differences in slopes may reasonably be distinguished from the noise; moreover, any outlier can dominate the slope. If the real differences in slopes are as large as those generated in Burstein and Linn (1976), then it is important to take them into account. Whether the differences in real classrooms are of similar magnitude is somewhat problematic at this stage, however.

Since students within a classroom are not a random sample, but possibly are better thought of as fixed once the classroom is chosen, it is not clear how best to investigate the relative magnitude of signal and noise in the differences among within-classroom slopes. Linn and Burstein (1977) found little support for the notion that slopes varied systematically when a posttest in reading was regressed on a pretest in reading (Figure 1)

and only limited support for the notion based on a similar set of regressions for math using small samples of classrooms from the ETS BTES study (McDonald and Elias, 1976). But these analyses were based on traditional confidence intervals which treated each class as if the students in it were a random sample from a population. As already noted, the random sampling model is questionable in this situation.

While random sampling of students may not provide the best model, there is a need to allow for disturbances in the observed slope due to idiosyncratic occurrences at the time of measurement. Just as an individual's observed score is distinguished from an underlying true score in classical test theory, there is a need to distinguish between the observed measure for the group (in this case the slope) and an underlying "true" slope.

Several approaches can be used to investigate the relative size of signal and noise in the within-group slope estimates. Considence intervals can be computed for the within=group slopes for selected sets as was done by Linn and Burstein (1977) for BTES data. The Jacknife procedure (Mosteller, and Tukey, 1977) can also be used to estimate slopes and confidence.

Relations of Slopes to Other Group-Level Indices

If slopes are to provide a useful addition to the array of outcomes, they must be distinct from other indices. Linn and Burstein (1977) have investigated this property of slopes. For three separate data sets (BTES data on classrooms collected by ETS (McDonald and Elias, 1976); Michigan Assessment Data on schools reported in Marco (1974); and IEA data on schools (See Table 2); they found that though pretest and posttest means correlated with each other

in the range of .5-.8, the correlation of within-group slope with either means or standard deviations (or, for that matter, with skewness and kurtosis indices and sample size) are much lower and, except for the pretest standard deviations (which are spuriously related to slopes), are rarely significant.

The results cited suggest that slopes are sufficiently distinct from means and standard deviations to warrant further consideration.

Relation of Slopes to School and Class Characteristics

The final line in an investigation of the potential utility of within-group slopes involves their relationships to measures of school and classroom processes. Preliminary results of an analysis of science achievement data on U.S. 14-year-olds in the IEA study (Burstein, 1978) provided tantalizing evidence of the possible payoff from this activity. Burstein found that the within-school slopes of science achievement on a verbal ability measure (assessed concurrently) were significantly and positively related to school mean responses of pupils on indices of exposure to science instruction and of the degree to which pupils reported instructional practices which emphasized exploration--discovery methods of instruction. (See Table 3). These significant results occurred despite controls for pretest and posttest means and standard deviations and pupil home background measures.

The results described above fit in well with recent research on informal/open/individually-guided/unstructured instruction (see particularly Rosenshine (1978) and Stebbins and others (1977)). Instruction which emphasizes student self-direction (selection) of learning goals and methods tends to exacerbate pre-existing differences in pupil skills. Higher-ability

students tend to make more appropriate choices and achieve at a faster rate than lower-ability students.

The steeper within-group slopes with greater opportunities for exposure to instruction and with greater emphasis on individual exploration cited above are consistent with expectations from other research and suggest the need for similar investigations with other data sets.

Estimating Within-Group Dependency in Multilevel Analysis

The Problem of Dependency among Observations within Units

The problem of dependence among observations within groups is endemic to research on hierarchically nested school data, and can be especially critical when intact classrooms are investigated. Cronbach and Webb (Cronbach, 1976; Cronbach and Webb, 1975; Webb, 1977) have argued that when intact groups are assigned to instructional treatments, the students in those treatments cannot be considered independent units and therefore, the typical analyses based on all individuals pooled across groups can be justifiably criticized as inappropriate.

The crucial problem in ignoring group membership is that educational treatments are not administered independently to individuals (Wiley, 1970). Individuals within the classroom have shared experiences. This non-independence of individuals within the group can be expressed by an intraclass correlation structure. The consequences of ignoring this intraclass structure (i.e., treating individuals as independent by ignoring group membership) are serious (Walsh, 1947; Weibull, 1953).

Recent work by Glendening (1976) provides a thorough discussion of the problem in the experimental design frame of reference. (The work of Glass and Stanley (1970), and Peckham, Glass and Hopkins (1969) is also summarized by Glendening 1976.) Glendening simulated the effects of violating the assumption of independence within the context of a balanced two-level hierarchically-nested design, with subjects (S) nested within classrooms (C) and classrooms nested within treatments (T). She operationally defined independence as that condition wherein the expected mean square between classrooms, EMS(C:T), equals that within classrooms, EMS(S:CT). She found that a model with the pupil as the unit or a conditional model where preliminary test of independence is followed by a choice of unit of analysis for testing treatment effects, yielded spuriously small error terms and therefore, too liberal tests of treatment effects. Glendening concluded that the researcher must choose a priori between the class (dependence) or student (Independence) as the unit, but acknowledged the complications of obtaining prior knowledge about independence of response.

While Glendening and Porter focused on the implications on intraclass correlation for the analysis of experimental data, Webb (1977) was concerned with the antecedents of such intraclass relations in research on group process. Webb compared learning in interacting groups and learning singly, attempting to explain differences as a function of the characteristics of the individual, the group, and the group process. The group process results provided a key to understanding why some students learned best in interacting groups, whereas others did best learning singly. In general, group members who actively participated in discussions did better than those who did not actively participate, and did at least as well as after individual learning. Whether a pupil actively participated was

related to the pupil's ability ranking within the group and the range and level of ability in the group. Knowing the abilities of the students in a group, one could predict fairly well wh- interacted with whom and, consequently, who did best.

The results of this highly structured study suggest that knowledge of group processes in a particular class is crucial for understanding the degree to which students are workinb together--and therefore crucial for estimating degree of dependence in the class. Studying group process may be the only way to get at this dependence. Unless students in a class are receiving completely individualized instruction, rarely will it be tenable to base analyses on the assumption of an intraclass correlation of zero. Unless all students are receiving exactly the same instruction and interact with fellow students in the same manner and amounts of time, an intraclass correlation of one is unreasonable. Examination of lower-level processes will help locate the intraclass correlation on the continuum between 0 and 1.

Webb suggests that the above procedures may be generalized to real teacher-taught classrooms, considering interactions between teacher and students, interactions among students and characteristics of students (abilities, personality variables) and teachers. In the long run, one hopes to be able to predict student performance from a combination of these variables.

Clearly, research on most educational phenomena will involve dependent observations. Moreover, dependence cannot be viewed as an all-or none phenomenon--it is a matter of degree. It depends on what is being measured (the outcome) and the "treatments" or "causes" under study.

It is also a function of the composition of the units and the nature of the grouping mechanism as Webb (1977) has demonstrated. Therefore tests for independence and adjustments for intraclass correlations are more appealing than automatic aggregation to the classroom level.

Analytical methods are needed which will account for the degree of dependency and make adjustments, where appropriate, to the estimated effects and associated estimates of precision. Moreover, estimators of dependency may be useful as indicators of classroom process. That is, it may be possible to relate these estimated relationships to characteristics of students, teachers, and instructional context.

Concluding Remarks

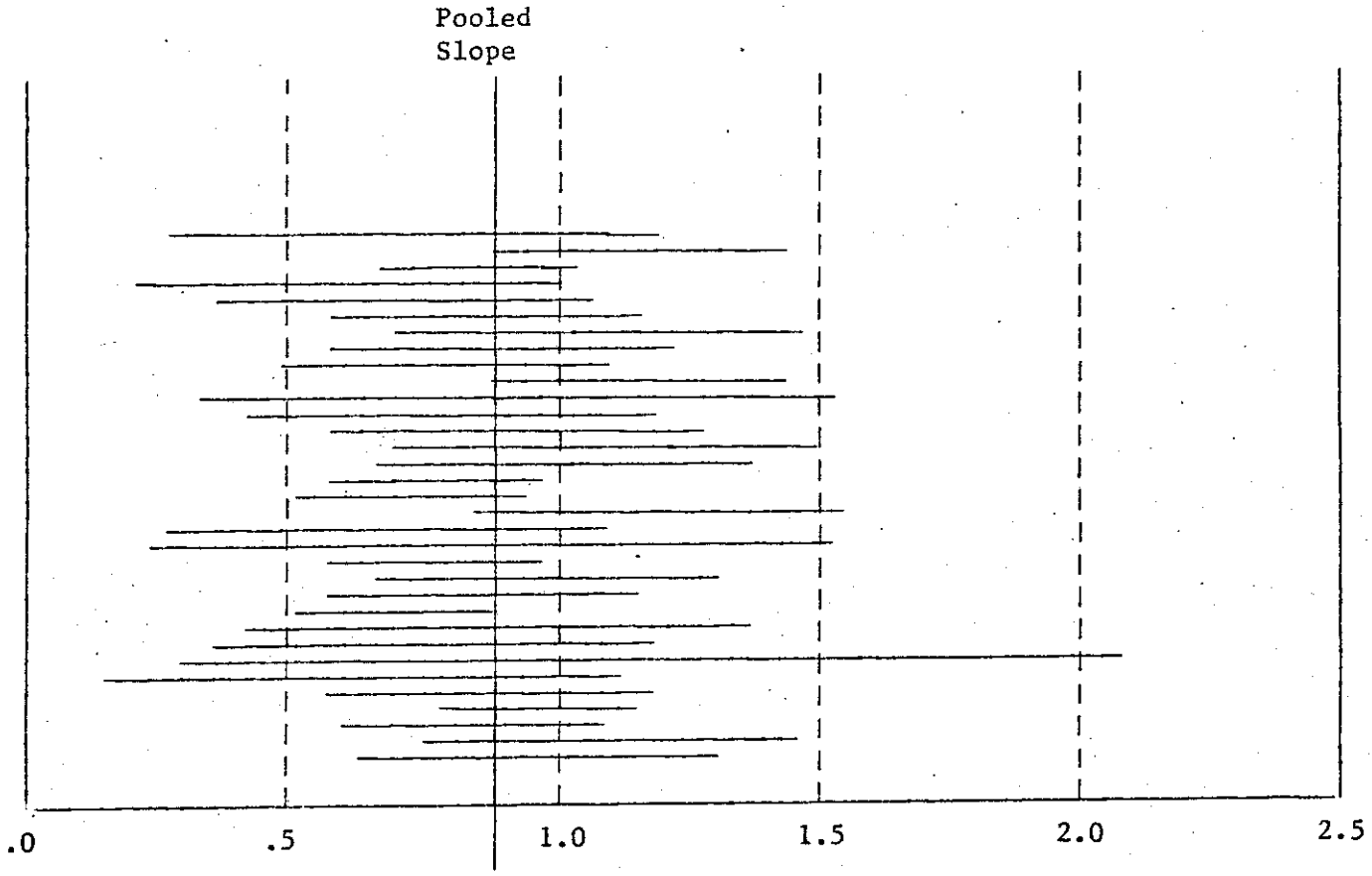
The topics discussed in this paper are a subset of a broader range of issues and problems which require more attention over the next few years. Table 4 lists a variety of types of studies and types of outcomes for which multilevel analysis issues must be resolved. It is unclear what form the final products of the investigation of the analysis of multilevel data will take, but it is possible to imagine the following scenario. As a preamble, we point to a trend developing in educational evaluation for the conduct of what Glass (1976) has termed "meta-analyses" (see also Light and Smith, 1971). Persons conducting meta-analyses seek to accumulate knowledge about the impact and characteristics of a particular educational innovation by aggregating findings across numerous investigations of the phenomena.

There would seem to be a natural parallel to meta-analysis which is relevant to the examination of alternative methodological approaches for the analysis of multilevel data. There are two key obstacles to the development of appropriate methodologies in this context. First, the

available methodological approaches vary greatly in the degree to which they are theory-based as opposed to ad hoc. Second, all currently available empirical data sets suffer from a variety of inadequacies which, taken singly, limit their utility for comparing alternative methodological approaches.

We believe that it is important to identify approaches which are practically viable as well as theoretically sound and which are usable with actual as well as hypothetical data. Therefore, we propose that in addition to the studies of the variation in analytical properties across approaches with hypothetical data, the alternative approaches should be applied to a wide variety and sizeable number of actual data sets, each with a potentially differing set of inadequacies. In this way we hope to learn more about both the methods (e.g., which are more generally usable; which behave similarly for specific kinds of data sets) and the influence of data limitations on methods (e.g., the exclusion of what types of information makes different approaches impractical or unattractive).

FIGURE 1



Ninety Percent Confidence Levels for Regressions of Posttest in Reading on Pretest in Reading for 33 Second-Grade Classrooms (Data from McDonald & Elias, 1976)

Source: Linn & Burstein, 1977

Table 1. Between-student, between-school[a] and pooled within-school regression analyses of factors affecting science achievement (RSCI) for 14-year-olds from the IEA study in the United states

Metric Regression Coefficients

Variable	Between School		Pooled Within-School		Between Student	
	United States	Sweden	United States	Sweden	United States	Sweden
Sex	-6.620 (3.90)[b]	-3.362+ (1.68)	-3.853* (11.30)	-5.281 (14.18)	-4.157 (11.87)	-5.165 (13.68)
Work Knowledge	.876 (6.47)	.569+ (3.24)	.812 (21.53)	.773 (18.36)	.861 (23.22)	.754 (18.02)
Father's Occupation	.843 (2.87)	.194+ (.67)	.307* (3.91)	.297 (3.64)	.487 (6.39)	.256 (3.26)
Number of Books in Home	3.577 (3.37)	1.217+ (.92)	1.223* (5.36)	1.324 (5.04)	1.661 (7.20)	1.324 (4.99)
Grade	1.390 (1.69)	4.186+ (2.49)	2.291* (5.25)	2,941* (7.25)	1.912 (5.24)	3.083 (7.65)
Science Study	.110 (2.34)	-.149+ (2.32)	.065* (3.90)	-.122* (5.70)	.066 (4.29)	-.125 (6.16)
Exploratory Methods	.240 (1.50)	.382+ (1.10)	.067* (1.25)	-.099* (1.31)	.130 (2.52)	-.064 (.85)
R^2	72	31	31	34	39	34
Number of Schools	107	93				
Number of Students	1806	1675				

[a] The between-school analyses are run with each school weighted by the numbers of students. However, all t-statistics were adjusted to reflect the number of schools rather than the number of students.

[b] t-statistics are reported in parentheses.

+ variable for which between country differences were significant at $p<.05$.

*Within-country variables for which the between school and with-school coefficients differ by at least two standard errors.

Table 2. Correlations among descriptive statistics from IEA data for the United States (N = 107).

		MATH						
		β	$\bar{Y}$	$\bar{X}$	S_y	S_x	N	Mean
READING	β		.10	.14	.59*	.13	.29*	.85
	$\bar{Y}$	-.02		.88*	-.09	-.24	.27	177.74
	$\bar{X}$	.09	.93*		-.12	-.34*	.20	152.50
	S_y	.48*	-.18	-.21		.78*	-.38*	18.31
	S_x	-.04	.12	.02	.65*		.31*	18.41
	N	-.35*	.23	.19	-.19	-.05		
	Mean	.90	173.20	156.27	34.03	31.78		

SOURCE: Burstein, 1978.

Table 3. School-level regressions of means, standard deviations, and slopes on background and school characteristics for the United States Population II, IEA Study.

	Dependent Variables					
	Metric Coefficient			Standardized Coefficient		
Independent Variable	Science Mean	Science SD	Slope[a]	Science Mean	Science SD	Slope
Sex	-6.744 (-4.24)[b]	-2.443 (2.26)	-.608 (-2.43)	-.390	-.220	-.236
Word Knowledge	1.056 (8.33)	.109 (1.27)	.058 (2.94)	.640	.126	.282
Father's Occupation	.437 (1.50)	.306 (1.55)	.014 (.31)	.149	.153	.031
Number of Books in Home	4.513 (4.84)	1.574 (2.49)	-.295 (-.20)	.436	.241	-.020
Science Study	.074 (1.72)	.038 (1.29)	.021 (3.17)	.169	.128	.302
Exploratory Methods	.317 (1.94)	.226 (2.04)	.072 (2.81)	.191	.200	.270
R^2	.76	.34	.27			

[a] Slopes from within-school regressions of science score on word knowledge score

[b] t-statistics in parentheses

SOURCE: Burstein and Miller, 1978

Table 4. Classifications of types of studies and types of outcomes for the investigation of educational effects.

1. TYPE OF STUDY

   A. MANIPULATION

      1. EXPERIMENTAL/TRUE -- "Units" assigned to alternative treatments or Treatment/Non-treatments; some form of manipulation

         a. Random Assignment of Pupils from Classrooms to Treatments -- Pupils randomly assigned to treatment conditions; treatment outside of normal class routine; treatment non-group work

         b. Random Assignment of Pupils from Classrooms to Groups -- Pupils randomly assigned to treatment groups; treatment outside normal class routine

         c. Random Assignment of Pupils to Classes -- Pupils randomly assigned to classes; classes randomly assigned to treatments

         d. Random Assignment of Partial Classes to Treatments -- Portions of class randomly assigned to different treatment conditions.

         e. Random Assignment of Intact Classes to Treatments -- Students assigned to classes on unknown non-random basis; intact classes assigned to treatments

      2. EXPERIMENTAL/ATI -- Conditions under I with additional question of interaction with entering characteristics

      3. EXPERIMENTAL/LONGITUDINAL -- Repeated measurement (mastery testing, sequential analysis of behavior and interaction patterns, persistence) in context of empirical studies

   B. NON-MANIPULATION

      1. NON-EXPERIMENTAL/CROSS-SECTIONAL -- Large-scale cross-sectional survey of pupils, teachers/classrooms, schools, etc. for purpose of establishing educational school/teacher effects model.

      2. NON-EXPERIMENTAL/LONGITUDINAL -- Large-scale longitudinal survey (e.g., income maintenance, voucher study, Follow Through Evaluation).

      3. NON-EXPERIMENTAL/OUTLIER (RESIDUAL) ANALYSIS -- Develop indices of effects of system over and beyond what can be anticipated by entering characteristics

Table 4 Continued

4. NON-EXPERIMENTAL/CONTEXTUAL (COMPOSITIONAL) EFFECT -- Examination of whether the composition/frog-pond/normative climate of institution has an effect.

II. TYPE OF OUTCOME

A. SHORT TERM -- Duration of a lesson to, say, a year

1. Specific Cognitive Objective -- Single content domain/objective in an instructional sequence

2. General Cognitive Objective -- Standardized achievement test or total score over multiple objectives CRM

3. Affective Objective -- Attitude toward self and subject matter, efficiency

4. Group Behavior -- Peer socialization, group cohesiveness, group interaction

B. LONG TERM -- Duration of multiple years, retrospective academic antecedents

1. General Cognitive Outcome -- Standardized test or cumulative grades (e.g., SAT as outcome prediction of future grades from earlier test scores)

2. Educational Attainment -- Level of education

3. Occupational Attainment -- Level of occupation (social stratification theory)

4. Career Plans/Career Satisfaction

5. General Mental Health

Footnote

[1]An earlier version of this paper was presented at the Institute for Research on Teaching, Michigan State University, East Lansing, Michigan, December 10, 1977. This paper presents work partially supported by the National Institute of Education contract NIE G-78-0113 with the Center for the Study of Evaluation, University of California, Los Angeles, and by a grant from the Spencer Foundation to the Graduate School of Education, University of California, Los Angeles. The contents of the paper in no way reflect official opinions of the organizations mentioned above and they are not responsible for the interpretations made herein.

References

Bloom, B. S. Comments on E. E. Wiley, "The Design and Analysis of Evaluation Studies." In M. D. Wittrock and D. E. Wiley (Eds.), The Evaluation of Instruction: Issues and Problems, New York: Holt, Rinehart and Winston, 1970.

Brophy, J. E. The Student as the Unit of Analysis. Research Report No. 75-12, University of Texas at Austin, 1975.

Brown, W. and Saks, D. H. "The Production and Distribution of Cognitive Skills within Schools." Journal of Political Economy, 1975, 83: 571-593.

Burstein, L. "Data Aggregation in Educational Research: Applications," Technical Report No. 1, Consortium on Methodology for Aggregating Data in Educational Research. Milwaukee, Wisc.: Vasquez Associates Ltd., March 1975.

Burstein, L. "Alternative Apporoaches for Assessing Differences Between Grouped and Individual-Level Regression Coefficients." Sociological Methods and Research, 1978, 7, 5-28

Burstein, L., Fischer, K., and Miller, M. D. "Social Policy and School Effects: A Cross-National Comparison." Paper presented at the Ninth World Congress of Sociology, Uppsala, Sweden, August, 1978.

Burstein, L., and Linn, R. L. "Detecting the Effects of Education in the Analysis of Multilevel Data: The Problem of Heterogeneous Within-Class Regressions," Paper presented at the Conference on Methodology for Aggregating Data in Educational Research, Stanford, CA, October, 1976.

Burstein, L., Linn, R. L., and Capell, F. "Analyzing Multilevel Data in the Presence of Heterogeneous Within-Calss Regressions." Journal of Educational Statistics, 1978, 3(4), 347-383.

Burstein, L., and Miller, M. D. Alternative Analytical Models for Identifying Educational Effects: Where are We? Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, March, 1978.

Burstein, L., and Smith, I. D. "Choosing the Appropriate Unit for Investigating School Effects," The Australian Journal of Education, 1977, 21, 65-79.

Cline, M. D., Ames, N., Anderson, R., Bale, R., Ferb, T., Joshi, M., Kane, M., Larson, J., Park, D., Proper, E., Stebgins, L., Stern, C. Education as Experimentation: Evaluation of the Follow Through Planned Variation Model. Vols. 1A; 1B. Cambridge, Mass.: Abt Associates, 1974.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, S., Weinfeld, F. D., and York, R. L. Equality of Educational Opportunity (Volumes 1 and 2). Washington, D. C.: U.S. Government Printing Office, 1966.

Comber, L. C., and Keeves, J. P. Science Education in Nineteen Countries: International Studies in Evaluation (Volume 1). Stockholm: Almqvist and Wiksell; and New York: Wiley, 1973.

Cronbach, L. J. (With assistance of J. E. Deken and N. Webb). "Research on Classrooms and Schools: Formulation of Questions, Design, and Analysis," Occasional Paper, Stanford Evaluation Consortium, July, 1976.

Cronbach, L. J. and Webb, N. "Between-Calss and Within-Calss Effects in a reported Aptitude X Treatment Interaction: Reanalysis of a Study by G. L. Anderson." Journal of Educational Psychology, 1975, 67: 717-724.

Duncan, O. D., Cuzzort, R. P., and Duncan, B. D. Statistical Geography: Problems in Analyzing Areal Data, Glencoe, Ill.: Free Press, 1961.

Feige, E. L., and Watts, H. W. "An Investigation of the Consequences of Partial Aggregation of Micro-Economic Data." Econometrica, 1972, 40: 343-360.

Glass, G. V. "Primary, Secondary, and Meta Analysis of Research," Educational Researcher, 1976, 5, 3-8.

Glass, G. V., and Stanley, J. C. Statistical Methods in Education and Psychology. Englewood Cliffs, N. J.: Prentice Hall, 1970.

Glendening, L. The Effects of Correlated Units of Analysis: Choosing the Appropriate Unit. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1976. (Also, unpublished Doctoral Dissertation, Michigan State University, East Lansing, Michigan, 1975.

Glendening, L., and Porter, A. C. Interdependence and Selecting the Appropriate Unit of Analysis (or the Appropriate Analytical Model). Technical Report No. 23, Consortium on Methodology for Aggregating Data in Eeucational Research. Milwaukee, Wisc.: Vasquez Associates, Ltd., October, 1976.

Haney, W. Units of Analysis Issues in the Evaluation of Project Follow Through. Unpublished Report. Cambridge, Mass.: Huron Institute, 1974.

Hannan, M. T., and Burstein, L. "Estimation from Grouped Observations." American Sociological Review, 1974, 39: 374-92

Hannan, M. T., Freeman, J., and Meyer, J. W. "Specification of Models of Organizational Effectiveness." American Sociological Review, 1976, 41: 136-43.

Hannan, M. T., and Young, A. "Estimation in Panel Models: Results on Pooling Cross-Sections and Time Series." In d. Heise (Ed.) Sociological Methodology, 1976-77. San Francisco: Jossey-Bass, 1977.

Keesling, J. W., and Wiley, D. E. "Regression Models for Hierarchical Data," Paper presented at the Annual Meeting of the Psychometric Society, Stanford University, 1974.

Klitgaard, R. E. "Going Beyond the Mean in Educational Evaluation," Public Policy, 1975, 23: 59-79.

Light, R. J., and Smith, P. V. "Accumulating Evidence: Procedures for Resolving Contradictions Among Different Research Studies." Harvard Educational Review, 1971, 41(4), 429-471.

Linn, R. L., and Burstein, L. Descriptors of Aggregates. CSE Technical Report Series, Los Angeles: Center for Study of Evaluation, 1977.

Lohnes, P. "Statistical Descriptors of School Calsses," American Educational Research Journal, 1972, 9: 574-556.

Marco, G. L. "A Comparison of Selected Effectiveness Measures Based on Longitudinal Data," Journal of Education and Measurement, 1974, 11, 225-234.

McDonald, F. J., and Elias, P. J. The Effects of Teaching Performance on Pupil Learning, BTES Phase II Final Report, Educational Testing Service, 1976.

Mosteller, F., and Tukey, J. W. Data Analysis and Regression: A Second Course in Statistics, Reading, Mass: Addison-Wesley, 1977.

Peckham, P. D., Glass, G. V., and Hopkins, K. D. "The Experimental Unit in Statistical Analysis." Journal of Special Education, 3: 337-349, 1969.

Rock, D. A., Baird, L. L., and Linn, R. L. "Interaction Between College Effects and Students' Aptitudes," American Educational Research Journal, 1970, 7: 109-121.

Rosenshine, B. V. "Formal and Informal Teaching Styles; A Review of S. N. Bennett and Others, 'Teaching Styles and Pupil Progress,'" American Educational Research Journal, 1978.

Snow, R. E. "Learning and Individual Difference." In L. S. Shulman (Ed.) Review of Research in Education, Volume 4. Itasca, Ill.: F. E. Peacock Publishers, 1976.

Stebbins, L. B., St. Pierre, R. G., Proper, E. C., Andersen, R. B., and Cerva, T. R. Education as Experimentation: A Planned Variation Model, Vol. IV-A, Cambridge, MA: Abt Associates, 1977.

Walsh, J. E. "Concerning the Effect of Intraclass Correlation on Certain Significance Tests." Annals of Mathematical Statistics, 1947, 18: 88-96.

Webb, N. Learning in Individual and Small Group Settings. Unpublished Doctoral Dissertation, Stanford University, 1977.

Weibull, M. "The Distribution of t- and f-statistics and of Correlations and Regression Coefficients in Stratified Samples from Normal Populations with Different Means." Skandinavisk Aktuarietidskrift, 1953, 36: 407-416.

Wiley, D. E. "Design and Analysis of Evaluation Studies." In M. D. Wittrock and D. E. Wiley (Eds.), The Evaluation of Instruction: Issues and Problems. New York: Holt, Rinehart, and Winston, 1970.

Wittrock, M. D., and Wiley, D. E. The Evaluation of Instruction: Issues and Problems. New York: Holt, Rinehart and Winston, 1970.