Analysis of Educational Effects From a Multilevel Perspective: Disentangling Between- and Within-Class Relationships in Mathematics Performance

Leigh Burstein
Robert L. Linn

CSE Report No. 172
1982

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The research reported herein was supported in part by a grant to the Center for the Study of Evaluation from the National Institute of Education, U. S. Department of Education. However, the opinions and findings expressed here do not necessarily reflect the position or policy of NIE and no official endorsement should be inferred.

TABLE OF CONTENTS

INTRODUCTION

This paper discusses how the multilevel character of data should influence analyses of the determinants of educational achievement. The discussion is grounded in the context of planning explanatory analyses for the Second International Mathematics Study (SIMS, Travers, 1980), a longitudinal cross-national investigation of instruction and achievement in mathematics classrooms. Our intent is to describe how a multilevel perspective in the specification of substantive questions and, as a consequence, in the analysis strategies employed, can potentially clarify the ways in which instructional practices affect student performance.

As we see it, a multilevel perspective implies a point of view in the examination of educational data. One begins with the obvious notion that the process of schooling takes place in a multilevel (more precisely, hierarchical) organization involving, in its most concise form, three levels: pupils, classroom/teachers, schools. Pupils receive instruction, either individually or in groups, from teachers in classrooms; these classrooms, and the pupils and teachers within them, are located within the schools.

Because of organizational features, the effects of schooling on individual pupil performance can exist both between and within the levels of the educational system. Moreover, analyses of data at different levels address different questions and, typically, research on schooling is focused on the questions relevant to the particular level of interest.

To state that educational processes are multilevel does not negate traditional concerns in research on educational effects, however. As is

usually the case, we are ultimately concerned with the outcomes of individual pupils (Averch et al., 1972). But, the choice of the performance of individual pupils as an overriding focus does little to resolve the question of how data on pupils within classrooms (within schools) should be analyzed. On the contrary, the analytical model should be dictated by the substantive processes under investigation (Burstein, Linn & Capell, 1978; Cronbach, 1976).

Within the domain of research on educational effects, the substantive processes in operation are functions of pupil characteristics (e.g., aptitude, previous exposure, motivation), classroom characteristics (e.g., teacher style, peer abilities and support, instructional content, organization, and atmosphere), and school characteristics (e.g., physical resources, atmosphere conducive for learning). Moreover, these substantive processes may have collective (for the class or school as a whole) as well as individual effects. Given these features of multilevel educational data, the primary difficulties lie in specifying the proper substantive questions and identifying from the multiple levels evidence that potentially leads to resolution of the questions.

In what follows we discuss how a multilevel analysis of the SIMS study data might proceed. We attempt to present both the general questions that the multilevel analysis can address and specific features of a possible analytical framework.

BASIC RESEARCH QUESTIONS

The overall purposes of the SIMS (Schwille, 1978; Travers, 1980)

emphasize the examination of the relationship of curriculum policies and classroom practices to student achievement. As we suggested earlier, the processes of interest are inherently multilevel and several features of the study design make it possible to disentangle these processes to some degree.

In the longitudinal mathematics study, students are matched with their teachers; there are both pre- and post measures of pupil performance; detailed information on instructional practices will be gathered, and data will be collected from multiple classrooms in each school. Under these conditions it is possible to decompose variation in individual pupil performance into between-school, between-classroom-within-school, and within-classroom components. Furthermore, depending on the source and quality of the information, it should also be possible to decompose variation in curriculum policies and instructional practices into similar components, though there is some question as to whether the instructional experiences of the individual student are adequately measured.

The patterns exhibited in the decompositions are likely to be of interest to curriculum planners and classroom researchers if previous empirical investigations of this type (e.g., Wiley & Bock, 1967; Murname, 1975; Rakow, Airasian & Madaus, 1978) are any indication of the usefulness of the technique. For example, if variation among classrooms in content exposure for a given topic is large (as evidenced by a large between-class component), we might anticipate a shift in variation in pupil performance from within-class and between-school sources at the pretest to between-classrooms at the posttest. If, however, exposure

was determined by policies at the school rather than the classroom level, the shift in achievement variation would be toward a large between-school component on the posttest.

The variance decomposition described above is a rudimentary form of multilevel analysis. More refined and purposeful analyses are desirable. These analyses explore the relationships among background, process, and outcome across and within the levels of the educational system. A multilevel approach to the data analysis breaks two fundamental questions in school research--namely,

1. How do instructional practices affect pupil outcomes?
2. How does classroom (school) composition influence the effects of instructional practices on pupil outcomes?

--into a series of more specific questions which address more directly the phenomenon of between-group (school, classroom) and within-group influences on pupil performance.

The actual analyses in a multilevel investigation can be viewed in two ways. On the one hand, a multilevel analysis can be simply an examination of a variety of measures of group-level outcomes. In such analyses, some group-level indices (e.g., spread) are considered to reflect within-group phenomena. On the other hand, a multilevel analysis is treated as a decomposition of the individual-level relationship of outcome and input into between-group and within-group components. The emphasis on specific components is then determined by the questions of most interest to the investigator. In the following sections, these two ap-

proaches to multilevel analysis are discussed.

For the time being, we restrict attention to two levels -- classroom and pupil. An important difference between SIMS and previous IEA surveys (e.g., Husen, 1967; Comber & Keeves, 1973) is the emphasis that SIMS gives to the classroom as the focus of investigation. We consider this focus very desirable. The classroom corresponds more clearly than the school to the treatment unit. If there is variability in classroom processes within a school, then relationships between process variables and student achievement are apt to be obscured when data are aggregated to the school level.

Classroom Level Outcome Measures

Given an interest in classroom-level analyses, a number of questions remain about the construction of classroom-level indices from individual-level measures and about the specification of analytical procedures. In classroom research of the type considered here, the primary purpose of the construction of group-level outcome measures should be to find measures which best reflect the influences of classroom processes. The three types of indices which perhaps best characterize the results of classroom processes are classroom means, measures of spread, and within-classroom slopes.

Classroom Means. Probably the most common, and possibly the most useful, approach to the construction of classroom measures from individual measures is to simply compute class means. The class mean is a natural summary index. A comparison of the class mean on a pottest with the cor-

responding mean (based on the same students) on the pretest provides a direct indication of the average improvement (or decrement) on the subskill in question. It is clear that having means higher at posttest than pretest is desirable, and that there would be an interest in identifying classroom process variables that are related to the magnitude of the posttest mean given the value of the pretest mean.

To greatly oversimplify the situation, we could imagine a regression equation of the form:

$$\bar{Y}_{.j} = b_0 + b_1\bar{X}_{.j} + b_2C_j + E_j,$$

where $\bar{Y}_{.j}$ is the mean posttest score for class j; $\bar{X}_{.j}$ is the mean pretest score for class j; C_j is a classroom process measure; the b's are regression weights; and E_j is the residual for class $_j$. In practice, the equations will obviously be considerably more complicated and include several categories of variables. Some of these needed expansions will be considered below, but for now it is convenient to consider this oversimplified version.

Ignoring issues of specification error and measurement error in the predictors, it is clear that we would be interested in finding process variables with substantial, as well as statistically significant coefficients, b_2. Larger values of b_2 are desirable, and the process measure would distinguish among classes with higher versus low values of $\bar{Y}_{.j}$ for given values of $\bar{X}$.

Measures of Spread. The classroom means are potentially useful summaries of individual student scores. As is true of any summary measure, however, they cannot capture the full detail that is contained in the individual level scores. Some of the additional detail may contain important

information about student achievement differences that vary as a function of classroom process variables. Consider, for example, the hypothetical pretest and posttest distributions for two classes of 25 students shown in Table 1. In terms of the means, there is no difference between the two classes. Both classes have pretest means of 3 and posttest means of 4. However, the classes are distinguishable in terms of other characteristics of their distributions. The pretest distributions are identical, but the posttest distribution of class 2 is more variable than that of class 1. Differences between the two hypothetical classrooms in Table 1 would be obscured in equation 1 or similar analyses which used only the means as the summary index for a class.

- - - - - - - - - -

Insert Table 1 here

- - - - - - - - - -

Of course, the mean is only one summary measure of the individual scores within a classroom. Other characteristics such as the spread or shape of the distribution, or the proportion of students above some minimum, are potentially informative. For the simple example in Table 1, standard deviations or variances would obviously provide a distinction between the two classes. So, too, would an index such as the proportion of students which scores of, say, 5 or above. The best indices to use in addition to the mean are not obvious. But, there is a need, as several people have suggested, to look beyond the mean. As Wiley (1970, p. 267) has argued, "the objects of instruction might well affect other characteristics of the unit than mean level achievement. They might, in fact,

Table 1

Hypothetical Test Results For Two Classes with Equal Pre-and Posttest Means, Equal Prestest Variances, but Unequal Posttest Variances

Test score	Pretest Frequencies Class 1	Class 2	Posttest Frequencies Class 1	Class 2
7	0	0	0	3
6	0	0	2	3
5	2	2	6	4
4	6	6	9	5
3	9	9	6	4
2	6	6	2	3
1	2	2	0	3
0	0	0	0	0
Means	$\overline{X}._1 = 3$	$\overline{X}._2 = 3$	$\overline{Y}._1 = 4$	$\overline{Y}._2 = 4$
Standard Deviations	1.08	1.08	1.08	2.17

affect the distribution of achievement in the collectivity. If this is true, the moments of the achievement distribution might be used as criterion measures."

It is easy to imagine substantive contexts for which distributional properties other than the mean would be of concern. In schools and classrooms with a high proportion of children with poor entering performance, say in the bottom quartile, an effective school or teacher might be one which shifted a significant proportion of the pupils above the bottom quartile when instructional outcomes are measured.

A different thrust is suggested by the present emphasis on individualized instruction (cf., e.g., Bloom, 1976) in the schools. Theorists studying these models of instruction expect that instruction which more nearly meets the needs of the individual pupil will result in a high level of mastery and thereby little variability in outcomes. In practice, reports from developers of individualized instructional programs indicate that on general measures of cognitive outcomes, whether based on norm-referenced or criterion-referenced tests, pupils in individualized programs tend to show gains in mean performance comparable to those of pupils receiving traditional group instruction, but exhibit substantial increases in the spread of gains. Obviously, the variability of performance is of interest.

The idea of using distributional characteristics in addition to the mean as criterion measures has been reinforced by Lohnes (1972) and more recently by Klitgaard (1975). Using Michigan Assessment data, Klitgaard found that school standard deviations of student achievement were only

moderately related to background factors. This result stood in stark contrast to results using school means. One of the hypotheses that Klitgaard suggested for explaining the relatively low correlations between background factors and standard deviations of achievement is that schools have more of an effect on spreads than on means. He also recognized, however, that the smaller correlation might be the result of greater random fluctuation relative to the size of the real differences in the standard deviation or other statistical problems.

Brown and Saks (1975) also used the Michigan data and regressed school standard deviations on some unspecified background characteristics, and on school characteristics (average experience of teachers, student-teacher and professional staff ratios, and percentage of teachers with masters degrees). One or more of these school characteristics were found to have significant regression weights in analyses of three subpopulations of Michigan schools. Brown and Saks interpreted this result as an indication of the importance of considering distributional characteristics in addition to the mean.

Data for U.S. schools in the IEA Science Education Study (Comber & Keeves, 1973) provide another illustration of the relationships among alternate descriptive measures of aggregates. In this illustration, the groups are schools rather than classrooms and the data are cross-sectional rather than longitudinal. For the results reported in Table 2, the Word Knowledge Test (RWK) was treated as a proxy for the pretest measure and the Science Achievement Test (RSCI) was treated as the posttest. The correlation between school means on RWK and RSCI (.77) is lower than the

correlations between pretest and posttest means often observed, but still substantial. The within-school standard deviations on the two measures have a statistically significant correlation (.34) with each other. There are also substantial correlations between the school means and school standard deviations.

- - - - - - - - - -

Insert Table 2 here

- - - - - - - - - -

Of potentially greater interest than the correlations in Table 2 are those reported in Table 3. The correlations of school means and standard deviations for the RWK and RSCI tests with two indices of home background and two indices of school characteristics are reported in Table 3. Of particular interest are the correlations of EXPLORE with the standard deviations of RWK and RSCI (.24 and .32, respectively). EXPLORE is derived from school mean responses of students to questionnaire items that ask the degree to which instructional practices at a school emphasize exploration, i.e., discovery methods of instruction. Thus, there is a statistically significant correlation between the reported extent of exploration and the dispersion of the students' test scores within a school.

If the school standard deviations of RSCI are regressed on school standard deviations of RWK and the school EXPLORE index, the standardized partial regression weights are .28 for RWK SD and .25 for EXPLORE. Both values are larger than twice their standard errors. In other words, schools with greater emphasis on discovery methods tend to have larger dispersions of scores on the science test even when the within-school dispersion on the Word Knowledge test is held constant.

Table 2

Intercorrelations and Means of School Descriptive Statistics[1]

Descriptive Statistics	Descriptive Statistics[2]			
	RWK Mean	RSCI Mean	RWK SD	RSCI SD
RWK Mean	1.00			
RSCI Mean	.77*	1.00		
RWK SD	-.15	.19	1.00	
RSCI SD	.40*	-.60*	.34*	1.00
Grand Means	27.54	33.47	4.45	8.08

[1]Based on data for U.S. schools in Population II of the IEA Science Education Study (Comber & Keeves, 1973). N = 107 schools.

[2]The variables are: RWK - Raw Score on Word Knowledge Test, RSCI - Raw Score on Science Achievement Test.

*Significantly different than zero with p = .05.

- - - - - - - - - -

Insert Table 3 here

- - - - - - - - - -

The above examples are somewhat removed from the mathematics study in several respects. They involved the school rather than the classroom as the unit of aggregation. They were based on cross-sectional rather than longitudinal data. The measures of achievement were more broadly defined than the core subtests used in the mathematics study.

The number of students with pretest and posttest scores in a single classroom will often be fairly small. Consequently, the within-classroom standard deviations might be expected to be insufficiently stable to be very useful measures. But data from the first phase of the Beginning Teacher Evaluation Study (BTES; McDonald & Elias, 1976) suggest that standard deviations of individual classrooms are relatively stable over time. McDonald and Elias reported means and standard deviations for 33 second grade classrooms on a reading test and an overlapping set of 37 classrooms on a mathematics test. The number of children per classroom varied from 10 to 27.

The posttest standard deviations ranged from 38.80 to 65.85 in reading and from 18.17 to 42.92 in mathematics. In both cases, however, there was a strong positive relationship between the size of classroom standard deviation on the pretest and the corresponding value on the posttest. The correlation between pretest and posttest standard deviations was .65 for reading and .78 for mathematics (see Table 4). These are less than is typical of the correlations between pretest and posttest classroom means (.93 for reading and .88 for mathematics in the McDonald et al. Study), but still quite substantial. These correlations seem even larger when the

Table 3

Correlations of School Means and Standard Deviations With Indices of Home Background and School Characteristics[1]

	Test Data[3]			
	RWK-Mean	RSCI-Mean	RWK-SD	RSCI-SD
Home Background[2]				
POPOCC	.53*	.64*	.14	.41*
BOKHOM	.58*	.72*	.17	.50*
School Characteristics[4]				
SCISTUDY	-.08	-.03	-.11	-.01
EXPLORE	.17	.29*	.24*	.32*

[1]Based on data for U.S. Schools in Population II of the IEA Science Education Study (Comber & Keeves, 1973). N = 107 schools.

[2]POPOCC and BOKHOM are school means on measures of father's occupation and number of books in the home respectively

[3]See Footnote 2 of Table 2.

[4]SCISTUDY is an index of the amount of exposure students have to science and EXPLORE is an index of the degree of emphasis on exploration in the instructional practices at a school.

*Significantly different than zero with p = .05.

size of the sampling variability of standard deviations based on samples of from 10 to 27 is considered.

- - - - - - - - - -

Insert Table 4 here

- - - - - - - - - -

These correlations are high enough to suggest that the standard deviation is a sufficiently stable measure of the within-classroom spread of scores to be a potentially useful descriptor of classroom variability in achievement on pre-and posttests. Moreover, as can be seen in Table 4, the correlations of posttest standard deviations with posttest means (-.18 in reading, -.09 in mathematics) are small. Clearly, the standard deviations are providing stable information that is not redundant with the class means in this example. The test of the utility of measures of spread, however, ultimately depends on finding stable relationships of posttest spreads and other measures of student characteristics. It is toward this end that we suggest that within-classroom standard deviations be used as classroom measures to be investigated in regression and structural equation models. That is, at the overly simplified level of equation 1, an analogous equation would be used where standard deviations replace the pretest and posttest means.

Within-Classroom Slopes. Still staying with the classroom as the unit of interest, there are characteristics other than statistics involving only the distribution of the classroom posttest scores that might be of interest. If there is an interest in finding process variables that are related to big gains in performance for initially low achieving students, then the focus needs to be on that category of students. The SIMS

Table 4

Intercorrelations and Means of Classroom Descriptive Statistics with Math Test Results above the Diagonal and Reading Test Results Below the Diagonal[1]

Descriptive Statistics	Descriptive Statistics: Pretest Mean	Posttest Mean	Pretest SD	Posttest SD	Math Grand Means
Pretest Mean		.88*	-.34*	-.12	152.50
Posttest Mean	.93*		-.24	-.09	177.74
Pretest SD	.02	.12		.78*	18.41
Posttest SD	-.21	-.18	.65*		18.31
Reading Grand Means	156.27	173.20	31.78	34.03	

[1]Based on McDonald & Elias, 1976. The number of classrooms equals 33 for reading and 37 for math.

*Significantly different than zero with p = .05.

mathematics achievement measure for fourteen year-olds (Population A) included the same core test of eight items from each of five content areas (algebra, common and decimal fractions, geometry, measurement, ratio and proportion) given at both pretest and posttest, plus four additional and roughly parallel forms taken on a rotated fashion by students at the posttest. With the eight-item core subtest, one could imagine a focus on students at any particular pretest score. The posttest score for only those students would then be used to define a descriptive measure of the classroom to be related to other variables. For example, the conditional mean posttest score might be computed for each classroom with one or more students who had a selected pretest score. In this way, as many classroom indices could be computed as there were one or more students in a classroom with that particular pretest score.

Although it is intuitively reasonable to compare conditional mean posttest scores of different classrooms at each possible pretest level, it is awkward and not apt to be very fruitful for several reasons. It would involve as many indices of a classroom as there are distinct scores on the pretest. Many classrooms would not have any students with a particular pretest score and, therefore, would not have a conditional mean for that pretest value. Even where a classroom included one or more students with a particular pretest score, the number would often be very small and, hence, the conditional mean would be rather erratic.

Assuming that the conditional means are reasonably regular, then the potential information that they might provide could usefully be summarized by descriptive statistics describing the regression of posttest on pretest within a classroom. If it is further assumed that the regression is linear,

then the slope and intercept of the within-class regression would provide the desired information. Since information about overall level has already been considered in discussion of analyses of classroom means, our focus at this point is on the slope of the within-classroom regression line. The idea that within-group regressions may be of substantive interest was suggested earlier by Wiley (1970, p. 268).

Differences in the within-classroom slopes might be hypothesized to result from differences in classroom process. For example, it might be hypothesized that teachers (or classrooms) who are equally effective in terms of means might vary in terms of slopes, with an egalitarian or compensatory classroom producing a flatter slope and a meritocratic one producing a steeper slope (Burstein, Linn, & Capell, 1978).

The methodological literature on the study of aptitude-treatment interactions provides conceptual arguments for the examination of differences in within-classroom regressions. Snow (1977) reported data from a dissertation by Greene (1980) to demonstrate why specific within-class regression warrants attention. Greene's study investigated the effects of student choice and no choice about the organization of a series of booklet lessons on outcomes with general ability as the relevant aptitude measure. Each treatment group consisted of randomly assigned half-classes; total sample size was 165 students.

Based on an inspection of plots of within-half class regressions, Snow concluded that "... the impression of within-class heterogeneity . . . is striking" (Snow, 1977, p. 33). The slopes within either treatment group range from slightly negative to highly positive. Moreover, the slopes for different treatment groups within the same class exhibit a

similar degree of heterogeneity.

Data from the IEA Science Study (Comber & Keeves, 1973) for U.S. schools provide an illustration of the use of the slope of a within-group regression line as a descriptive measure of the group. The mean within-school slope from the regression of RSCI on RWK was .90 for 107 schools from Population II of the IEA Science Education Study (Comber & Keeves, 1973). There was substantial school-to-school variability in the within-school slopes. Over the 107 schools, the standard deviation of the within-school slopes was .45. The within-school slopes also had statistically significant positive correlations with the school means on RWK and RSCI and with the within-school standard deviation on RSCI (see Table 5).

- - - - - - - - - -

Insert Table 5 here

- - - - - - - - - -

The correlation of the within-school slope with within-school standard deviation on the dependent variable is to be expected since the slope is equal to the standard deviation of the dependent variable times the correlation between it and the predictor divided by the standard deviation of the predictor. Hence, the correlation of slope with the dependent variable standard deviation might be dismissed as an artifact. The correlations between within-school slope and the school means on RWK and RSCI, however, are not so readily dismissed. There is no necessary relationship between mean level and slope. The fact that schools with higher means tend to have steeper slopes is a finding of potential substantive interest (though such a result might possibly reflect floor and ceiling effects as well).

The within-school slopes were also correlated with the school means

Table 5

Correlations of Within-School Slopes With Other Within-School Descriptive Statistics[1]

Descriptive Statistic	Correlation
RWK Mean	.35*
RSCI Mean	.35*
RWK Standard Deviation	-.11
RSCI Standard Deviation	-.55*

[1]Based on data for U.S. schools in Population II of the IEA Science Education Study (Comber & Keeves, 1973). N = 107. See Table 2 for a description of the variables.

*Significantly different than zero, p = .05.

on the two measures of home background and two measures of school characteristics which were discussed in the previous section. These correlations are listed in Table 6. As is indicated in Table 6, within-school slope has significant positive correlations with BOKHOM and EXPLORE.

- - - - - - - - - -

Insert Table 6 here

- - - - - - - - - -

Although of modest magnitude (.27), the correlation between slope and EXPLORE is of special interest. It indicates that schools where more emphasis is placed on discovery methods of instruction tend to have steeper slopes than schools placing less emphasis on discovery methods. In other words, there is a tendency for greater exploration to be associated with larger individual differences in science scores compared to differences in word knowledge. These results are consistent with expectations from other research (Bennett, 1976; Rosenshine, 1978; Sorenson & Hallinan, 1977; and Stebbins et al., 1977). While they are not based on longitudinal data nor are they at the classroom level, they do provide an illustration of the notion of using slopes as outcomes and suggest that it would be desirable to use within-classroom slopes as indices in SIMS.

Multilevel Decomposition

In the previous section, a multilevel analysis approach was described in which the interrelations and antecedents of a variety of indices of group-level outcomes are investigated. It is perhaps preferable to approach the analysis of multilevel data more formally by modeling the decomposition of the data into components suggested by the kinds of educa-

Table 6

Correlations of Within-School Slopes with Indices of Home Background and School Characteristics[1]

Variable Type	Index	Correlation
Home Background	POPCOC	.18
	BOKHOM	.25*
School Characteristics	SCISTUDY	.19
	EXPLORE	.27*

[1]Based on data for U.S. schools in Population II of the IEA Science Education Study (Comber & Keeves, 1973). N = 107. See Table 3 for description of variables.

*Significantly different than zero with p = .05.

tional processes at work in such situations. Such a multilevel decomposition actually incorporates all the features of the previously described analysis, but couches them within a larger analytical model, which we describe below.

In a two-level (classroom, pupil) analysis, a multilevel decomposition of the factors influencing cognitive outcomes between and within classrooms would attempt to answer the following questions:

1. What is the relationship between the class composition at pretest and class-level performance at posttest? (Between-Class)
2. Given classrooms with similar entering composition, why is the posttest performance for some higher (more or less variable) than for others? (Adjusted Between-Class)
3. Why do some students in classrooms learn more than others given their relative differences in background? (Pooled Within-Class)
4. Why does the relationship of pretest (or any other background characteristic) to posttest vary across classrooms? (Specific Within-Class)

Each of the four questions deals with a unique source of information from a decomposition of the relationship of pupil background (denoted hereafter by X_{ij} for student i in class j) to pupil outcome Y_{ij} into between-class (between-class slope, adjusted between-class effects) and within-class (pooled within-class slope and specific within-class slopes) components.

This decomposition can be written as

$$
\begin{aligned}
Y_{ij} = Y.. &+ b_b(\overline{X}_{.j} - \overline{X}..) && \text{predicted between-class} \\
&+ (\overline{Y}_{.j} - \overline{Y}..) - b_b(\overline{X}_{.j} - \overline{X}..) && \text{adjusted between-class effect} \\
&+ b_w(X_{ij} - \overline{X}_{.j}) && \text{pooled within-class} \\
&+ (b_j - b_w)(X_{ij} - \overline{X}_{.j}) && \text{specific within-class} \\
&+ u_{ij} && \text{specific residual associated} \\
& && \text{with person j}
\end{aligned}
$$

In the above equation, b_b is the between-class slope from the regression of $\overline{Y}_{.j}$ on $\overline{X}_{.j}$; b_w is the pooled within-class slope from the regression of $(Y_{ij} - \overline{Y}_{.j})$ on $(X_{ij} - \overline{X}_{.j})$ across all classrooms; and b_j is the specific within-class slope from the regression of Y_{ij} on X_{ij} within the jth classroom.

Each of the four components of the multilevel decomposition can reflect the influences of instructional processes (between-classes, within-classes, or both). These components have been discussed in detail elsewhere (Burstein, 1980).

Between-Class Slope. The examination of the relationship of mean pretest to mean posttest should be standard for any longitudinal study of the educational performance of intact classrooms. The magnitude of the between-class slope provides an indication of the extent to which initial difference in the mean performance levels in a set of classrooms is maintained, exaggerated, or reduced at a later measurement occasion. If positive, as it typically would be, the slope reflects the tendency for classrooms with highest mean inputs to have high mean outcomes. A weak relationship implies little dependence of class mean outcomes on average initial performance.

Instructional practices do not directly affect the estimation procedure for the between-class relationship of pretest and posttest. However, their influence is evident when the investigation focuses on specific content areas. At the level of the specific subtest, the pre-post relationship will vary according to both instructional emphasis and the quality of instruction on the specific topic. This was clearly the case in Phase IIIB of the Beginning Teacher Evaluation Study (BTES; Fisher et al., 1978) where, for example, the smallest relationships between pretest and posttest performance in fifth grade mathematics occurred for the fractions subtest. Fractions was the topic which exhibited the greatest amount of coverage and the greatest variation in coverage. Moreover, there are some indications that instructional approach was also highly variable.

Clearly, an examination of the variation in the between-class relationship of input to outcome across content areas (and for that matter, across countries) would be of interest in the mathematics study. Such an examination would offer evidence as to which areas of mathematics instruction require more attention to achieve a desired relationship between entering status and outcomes at the class level.

Adjusted Between-Class Effects. In cross-sectional studies, the investigation typically concentrates on the mean posttest scores across classrooms. By measuring entering status, it is possible to adjust mean outcome differences for differences in mean inputs. These adjusted mean outcome differences, or adjusted between-class effects, can reflect the impact of the specific teacher/class (e.g., quality of classroom practices, content,

exposure, class cohesiveness) on the mean outcome of its pupils after controlling for the relationships of mean inputs to mean outcomes. Large positive adjusted classroom effects may be an indication of exemplary educational practices. The analysis should attempt to identify generalizable characteristics of the classrooms achieving large effects.

The actual precedures for identifying and investigating adjusted between-class effects in a multilevel decomposition mirror the analyses proposed for classroom means and measures of spread in the section on group-level outcomes. A set of more complicated class-level regression equations in the form of equation (1) are typically involved. In such equations, the coefficients associated with classroom process measures are interpreted to reflect the antecedents of adjusted classroom effects.

Pooled Within-Class Analysis. A pooled within-class analysis provides information about the tendency, across all classrooms, of students below the class average on the pretest to do better or worse on the posttest than the rest of the class. The within-class slope is simply a reflection of the relationship between relative standing within the group at two points in time.

An examination of pooled within-class slopes in the mathematics study would serve a purpose similar to an examination of the between-class slope. Variation in the pooled within-class relationships across content areas can be the result of the differential impact of instructional practices. (Floor or ceiling effects on the subtests are other plausible explanations.) If this were the case, comparisons of pooled within-class slopes of posttest performance in specific content areas on corresponding content-specific pretests and on indicators of ability prior to instruction (such as the

total score on the core pretest) could point to subject areas where instructional practices have exaggerated or reduced within-class differences in entry performance. More importantly, cross-national comparisons of pooled within-class slopes can potentially pinpoint the influences of national curriculum policies on the relative performance of students within classrooms.

Specific Within-Class Relationships. The final component to be considered involves more subtle and potentially treacherous terrain. In a typical study, the within-class relationship of entry performance to outcome (the within-class slope) varies across classrooms. On the one hand, this variation may be simply fluctuation due to small sample sizes, differences in within-class variability at pretest or posttest, or floor and ceiling effects. On the other hand, important differences in within-class instructional processes may be the source of this variation (Burstein et al., 1978).

The impact of specific classroom characteristics on the overall performance level of the class and on the relative performance of pupils within the class can be viewed in several ways. For example, in response to differing program requirements, school policies, instructional philosophies, and/or personal skills, teachers vary in their instructional goals and practices. These differences can result in varying outcomes, even when classes have similar distributions of entering performance.

The possibility of heterogeneous within-class slopes becomes obvious from the diversity of instructional practices that are characterized as being either individualized, compensatory, or traditional. Other things being equal, we expect that higher ability students may make more appro-

priate instructional decisions and learn at a faster rate than lower ability students in a class that emphasized individualized instruction with student self-pacing. Regardless of the overall effect on class mean performance, individualization could strengthen the relationship between entering ability and student outcome, thereby exaggerating pre-existing differences in pupil skills.

In contrast, another class might emphasize mastery learning, wherein it is expected that most students will master the curriculum content. Or, the teachers might compensate for preexisting differences by investing extra instructional resources on those students with poorest entering performance. These types of practices may or may not raise the mean performance of the class as a whole. It is likely, however, that such practices might reduce the relationship between entry performance and outcome, however; in other words, flatter slopes of the regression of output on input might be expected as the result of such compensatory practices.

The types of instructional strategies described above do exist in schools along with what might be called the more traditional pattern, with the same instructional program being delivered to all students in the class with heavy emphasis on whole-class activities (lecture, seatwork). While these different types of instructional strategies start with similar instructional resources, they may distribute these resources to individual students in a variety of ways. Thus, it is possible that varying instructional strategies would yield different class mean outcomes, and it is also plausible that different within-class slopes would result.

In the earlier section on group-level indices, we presented empirical evidence that within-class slopes can serve as outcomes that are distinct from other indices and are related to process measures. If the classrooms in the study reflect clear distinctions in instructional approach (e.g., teachers who emphasize the performance of computations with speed and accuracy vs. those who emphasize the logical structure of mathematics and the nature of proof), then we anticipate heterogeneous within-class relationships of the type found in the science study data. Therefore, while thorny theoretical problems and practical complications occur when slopes are used as outcomes, we think that the potential payoff is sufficient to warrant an attempt to examine within-class slopes in some detail in order to detect any pattern of relationships to content exposure and instructional approaches.

SPECIFICITY OF ACHIEVEMENT MEASURES

Before proceeding to the discussion of the general analytical model for the SIMS, we consider briefly the question of the specificity of the achievement measures to be used in the analysis. In the section on multilevel analysis approaches, the discussion focused on the eight-item core subtests as achievement measures. For reasons elaborated below, we feel that as far as is possible, the longitudinal analyses should be conducted with the core subtests (both content specific and process specific) as the outcomes to be explained.

To some extent, the rationale for our recommendation to emphasize the more specific core subtests rather than the composite tests reflects

a tradeoff between reliability and validity. On the one hand, it is quite likely that at the student level, the short measures of individual subskills will have modest reliability and relatively high intercorrelations compared to their reliabilities; the correlations at the classroom level are apt to be even higher. Furthermore, the more reliable composite measures offer practical advantages by reducing the volume of analyses.

Nonetheless, we think that it is important to keep the subtests separate for purposes of the major analyses. The premature formation of composite measures could obscure important relationships of classroom process variables to outcomes. Weak, but stable effects of process variables keyed to specific subskills could be diluted beyond detection when related to composite achievement measures.

The points above are simply an argument for the use of outcome measures that are more likely to be sensitive to instructional variation and therefore more valid in reflecting the effects of instruction. If, as anticipated, instructional activities and emphasis varies across classrooms, then analyses conducted with subtests as outcomes offer the most hope for identifying effects of classroom processes.

GENERAL PATH MODEL FOR MULTILEVEL ANALYSIS

To this point we have presented the features of a multilevel analysis as applied to data from studies such as the one currently being conducted. We have also stated a case for emphasis on content-specific and operation-specific subtests (or items) as outcomes. Now we turn briefly to the general structural model for mathematics performance on the specific subtests.

As in almost all larger-scale studies in research on learning in classroom and schools (e.g., Bloom, 1974; Carroll, 1963; Fisher et al., 1978; Harnischfeger & Wiley, 1976), the general model for pupil performance includes three general sets of variables: entry characteristics, instructional/schooling characteristics, and outcome status. The research problem is then to (a) measure adequately all components of each set of variables, (b) develop a justifiable theoretical model which explains the interrelations among specific variables within and among the three sets and (c) determine a suitable analytical strategy that will enable the investigator to estimate the relations in the posited model. Economists refer to this type of endeavor as the specification of the structural model.

The fact that we are discussing a multilevel approach to data analysis does not diminish the importance of proper model specification; in fact, if anything, the opposite is true. We still need to consider all three blocks of variables and to exercise care in ensuring that all key variables within the blocks are identified. But this latter task becomes more complex because of the need to consider not only whether a specific variable is important, but also at which level(s) its influences are likely to be felt.

The role of the cognitive pretests in the SIMS perhaps best illustrates the added concerns in a multilevel structural model. As things currently stand, the cognitive pretest has to serve several purposes in SIMS. The first function is to reflect current level of student knowledge in specific content areas. Second, as a whole, the student's pre-

test total score is the best available measure of entering mathematics ability. (This interpretation would, of course, be satisfactory only if students had but limited prior exposure to the topics on the pretest.) And if there is reason to believe that the effects of specific instructional practices depended on the student's math ability level, then the structural model would have to incorporate the cognitive pretest total and allow the examination of possible effects of interactions with instructional practices.

The third usage of the cognitive pretest is to create a suitable measure of mathematical abilities of the class as a whole. Individual student performance has typically been aggregated to the class level to describe the classroom composition in ability terms. This use and interpretation of the aggregated pretest has been hotly debated in recent years, largely because of a concern that its substantive means will be misinterpreted (Alwin, 1976; Burstein, 1980; Burstein et al., 1980; Farkas, 1974; Firebaugh, 1979; Hauser, 1971, 1974). Nonetheless, there are compelling reasons for arguing that class ability composition influences instructional practices which, in turn, affect the learning of individual students (Barr & Dreeben, in press; Bidwell & Kasarda, 1980; Brown & Saks, 1980; Webb, 1980). Thus, the structual model needs to incorporate some measure of this characteristic of the learning setting.

We do not intend to discuss each variable in the general model in as great a detail as we have done for the cognitive pretest. Instead, a table is provided (Table 7) delineating key variables (and their sources and descriptions) in the multilevel model, and an effort will be made to point out instances where a given variable might serve dual purposes (e.g.,

as a class-level aggregate or as a measure of the relative standing of a student within the class).

- - - - - - - - - - -

Insert Table 7 here

- - - - - - - - - - -

Entering Characteristics

Entering characteristics can play an essential role at multiple levels. We have already discussed the multilevel treatment of the cognitive pretest. Clearly, the fall ability status of the classroom can govern the level and the manner in which instruction proceeds. A student with a low pretest score in an otherwise high performing class can expect a much different educational experience than a similar student in a low performing class. The instructional emphasis and the teacher and peer pressures are likely to be highly dissimilar for the two students. The examination of the effects of within-class spread on the pretest and of the antecedents of within-class slopes are likely to provide some insight into the influence of ability composition.

Aggregate effects of other entering characteristics are possible, but less likely in a study with adequate pretest measurements. We anticipate that the typically large influence of aggregate measures of socioeconomic status on both aggregate and individual outcomes will be greatly reduced due to the inclusion of pretest scores. The kinds of community investment in educational experience and achievement orientation that aggregate SES typically reflects (Burstein, Fischer, & Miller, 1980) will already have influenced entering measures of prior exposure,

Table 7

Variables in the Multilevel Structural Model for Mathematics Performance

Variable	Source	Purpose/Comments
I. Entering Characteristics		
A. Ability		
1. General (Verbal reasoning)	None	Examines the possibility of high verbal loading of instruction or heavy demand on general reasoning ability.
2. Quantitative	Math pretest total (minus subtest of interest)	Measures quantitative ability relevant to to subject population.
B. Family Background		
1. SES	Student background questionnaire	Measure of adequacy of economic and cultural resources for support of schooling.
2. Home Encouragement	Student background questionnaire	Given equal resources, provision of atmosphere which encourages and supports academic performance.
C. Student Sex	Student background questionnaire	Of obvious concern, given common stereotypes regarding sex and mathematical performance.
D. Prior Exposure	Student prior Opportunity to learn	Topic coverage prior to pretest
E. Fall Status		
1. Cognitive pretest	Math subtest score at pretest	Self-explanatory
2. Attitude	Fall attitude scales	Motivation to learn mathematics during the school year.
3. Class composition	Fall test scores	Mean and range on pretest for the class on both total and subtest.

Table 7 (Cont.)

Variable	Source	Purpose/Comments
II. Instruction/Schooling		
A. Coverage		
1. Introduction	Teacher general questionnaires fall through spring	Newly introduced topics and amount of coverage.
2. Rehersal/Review	Teacher general questionnaires fall through spring	Amount of coverage of topic introduced in a prior year.
B. Quality		
1. Curricular approach	General and topic specific questionnaires	Teacher strategy for specific topics.
2. Organization/management	General student and teacher questionnnaires	Instructional grouping and pacing; settings.
3. Sequencing	Teacher general and topic questionnaires	Interrelation and ordering of topics.
4. Atmosphere	None	Clarity, enthusiasm, and warmth.
C. Resources		
1. Within-classroom	Teacher and student questionnaires	Calculators and other instructional materials and aids.
2. School-wide	School questionnaire	Math laboratory, computer, academic orientation.
III. Outcomes		
A. Cognitive Posttest	Spring subtest scores	---
B. Attitude	Spring attitude scales	Attitude towards math; toward further study of math.

performance, and attitude and thus represent more distal (and weaker) influences on outcomes once the latter are adequately measured.

At the classroom level, there may be effects of aggregate prior exposure and attitude. The former may be a more evident measure of fall status than the cognitive pretest, assuming that teachers are aware of the content coverage in prior years, but are less knowledgeable about specific student sknowledge. Aggregate attitude might be worth further examination to see if the overall class motivation to learn mathematics affects instructional practices and, in turn, individual and class-level performance and attitude on the posttest.

The above discussion emphasizes the aggregate effects of entering characteristics for the most part. In a multilevel analysis, one's relative position within the group is viewed as a possible influence worthy of investigation (Burstein et al., 1978; Davis, 1966; Firebaugh, 1979). The example of the low performing student in a high vs. a low class addresses this issue directly. Similarly, a student's previous instructional experiences and attitudes toward mathematics relative to the class can influence how well the student progresses and is treated by others. The data analysis had best be particularly sensitive to how the instruction responds to the presence of atypical students within classrooms and how these students in turn respond to the instruction and the instructional environment.

Instruction

It is clear that SIMS emphasis is on obtaining accurate information on instruction at the classroom level. This is a reasonable concern

given the curriculum orientation of the study. However, it is likely that class-level measures of instruction will miss any individualization of practices and exposure.

Little information on individual student instructional experiences is being collected in the SIMS study. Under these circumstances, it is unlikely that the pooled within-class analyses will include instructional process measures. Such measures, if available only at the class level, are constant for all students within the class and thus have no effect in a pooled within-class analysis. Given the above, we anticipate that instructional variables will exhibit their effects on mean outcomes or perhaps on spreads or within-class slopes. The influences of instructional grouping and pacing are likely to be especially evident when a variety of class-level outcome indices are examined.

Outcomes

The discussion on multilevel analysis approaches focused on the handling and interpretation of the outcome measures. Given the nature of the study, most of the analyses should be performed on class-level (class means, spreads, and slopes) and within-class (deviations from class mean) indices of pupil performance and attitude. Analyses involving pupil-level outcomes unadjusted for class membership would be potentially misleading. The decomposition of individual outcomes into between-class and within-class components as described earlier is more likely to resolve any confusion over the role particular instructional practices play in explaining pupil performance.

CONCLUDING COMMENTS

The same general issues arise in other large-scale investigations in classroom research. The key points can be summarized as follows: First, the questions of most interest have to do with the relationships between instructional practices and classroom composition and how these two sets of characteristics affect outcomes. Second, the analysis models we propose to use are largely regression-based methods applied to specific core subtests which in most cases would be decomposed into between-class and within-class effects.

The SIMS data are inherently multilevel and the instructional processes that are being examined have both between-class and within-class effects that warrant investigation. Moreover, the provision of multiple classrooms per school with longitudinal cognitive data and detailed instructional process information makes the SIMS particularly suited for a multilevel analysis.

Because of the likely interest in cross-national comparison, it may be desirable to run parallel analyses for each country with a uniform set of explanatory variables at each level of analysis. We have experienced some success with this latter approach in a multilevel analysis of science data from England, Sweden, and the U.S. (Burstein, Fischer, & Miller, 1980).

The real test of whether our questions and models are appropriate for a cross-national study lies in the process data to be collected. If there are cross-national differences in instructional approaches and content exposure, we can perhaps explain cross-national differences in cog-

nitive outcomes. This type of result is evident in the behavior of the EXPLORE scale from the IEA science study in the U.S. and in Sweden. In Sweden there was almost no emphasis on discovery approaches to science instruction and, consequently, little influence of EXPLORE on any measures of school outcomes. In contrast, the U.S. exhibited marked school-to-school variation in exploratory approaches and these school differences were related to differences in within-school slopes of science score on word knowledge score.

In addition, there are certain variables, such as conformity to national curriculum policies, whose influence can only be examined cross-nationally. The question of school-level and class-level conformity to national curriculum policies is certainly a topic of interest to curriculum planners.

The overall purpose of our proposed analysis should be to achieve a better understanding of the interrelationships among specific cognitive inputs and outcomes and strategies for mathematics teaching. We expect that as a general strategy, a careful multilevel examination of these interrelations which focuses on instruction and outcomes relevant to specific content areas has as much face validity as one can establish at present. Unless one is willing to argue that certain instructional approaches have uniform effects, regardless of the characteristics of the individuals receiving the instruction and the context in which the instruction is received, than an analysis that takes into consideration between- and within-class influences on performance would seem to be necessary.

Obviously, there are compromises reflected in the current study design that limit its compatibility with the questions and models we have

proposed. The major sources of problems have to do with the thoroughness of measures of entry characteristics and with the availability of corroborative data on classroom processes.

The limitations in the measurement of entry characteristics have more to do with what is excluded rather than the quality of available data. The entry measure that is most central to our analysis plans is the pretest. The next most important set of entry variables are a variety of aptitudes (e.g., verbal, reasoning, etc.) which might be activated (or inhibited) by specific instructional practices. There was no premeasurement of aptitudes other than the core test (which at best is a proxy for general mathematics ability) and this decision, though understandable in terms of time demands, is regrettable. The measurement of the other main elements of entry characteristics (SES, home encouragement, prior exposure, attitudes) seem adequate, especially once scaling problems are resolved.

The problems with the measurement of classroom processes are of two kinds: those which are unresolvable using strictly questionnaire data and those which are complicated by the type of questionnaire data gathered. Given other constraints, obtaining adequate observational data was not possible. On the other hand, improvements were feasible with the study's questionnaire on classroom processes. The final versions of the teacher general questionnaire and topic questionnaires show more promise than previous efforts. In fact, content exposure and the components of instructional quality as perceived by the teacher were measured as adequately as may be possible with questionnaire data.

The adequacy of the measurement of student perception of instructional experiences is less clear. Plans called for student reporting of content exposure in the same format as the teacher reports. More student reporting of other classroom practices as well would have been useful for at least two reasons. First, what the teacher tries to do and what the student perceives to be happening do not always agree. While it would be nice if it were the sent rather than the received message that counted, it is just as plausible to take the student's responses as fact as it is the teacher's. This is especially true when the questions involve low inference instructional behaviors (e.g., did you generally work in groups on projects; do you use calculators in class). Most questions from the teacher general questionnaire could have been asked of students as well. For example, the proportion of students responding positively to direct questions asking whether, during the last week of instruction, they were called upon in class, had their work checked by the teacher, and volunteered to answer a question may provide a better indication about these practices than the teacher's own recollection.

A second reason for urging the collection of individual student data on classroom practices is that this may be the only way to identify aberrant classes (large differences between aggregated student responses and teacher responses) or students (individual student responses which differ from the rest of the class). Either type of disagreement identifies unusual cases (classes or students). Isolating these cases could potentially sharpen interpretation from the remaining data. Furthermore, such cases may be of special substantive interest in their own right.

Once again, experiences from the Six Subjects Science Survey lend empirical support for the conceptual arguments for more extensive student reports. The EXPLORE scale which was based solely on student responses to specific questions about their instructional experiences behaved well in various secondary analyses (e.g., Burstein, 1980; Burstein, Fischer, & Miller, 1980; Kelly, 1978). For example, in reanalyses of U.S. science data, the EXPLORE scale was positively related to the slope from the within-school regression of science score on work knowledge score. Since there were precious few school characteristics with positive results in the science study, the successful use of the EXPLORE scale speaks well for similar attempts to use reports of low inference behaviors in the mathematics study.

The general issue of student reports warrant consideration. There was a great deal of resistance to anything that smacked of student evaluation of teacher. For that reason, we did not argue for direct student ratings of classroom climate or atmosphere, though this block of variables was included in our general path model (Table 7). Our sense is that classroom climate information is most useful for interpreting the behavior of those classrooms (and students) which behave atypically given similar entry characteristics and instructional experiences. Hopefully, if climate information is needed in this capacity, a measure can be constructed by the careful examination of a variety of indirect evidence from teacher and student questionnaires, while recognizing the high risk involved in overinterpreting such data.

REFERENCES

Alwin, D. F. Assessing school effects: Some identities. Sociology of Education, 1976, 49, 294-303.

Averch, H., Carroll, S. J., Donaldson, T., Kiesling, H. J., & Pincus, J. How effective is schooling? A critical review and synthesis of research findings (R-956-PCSF/RC). Santa Monica, Calif.: The RAND Corporation, 1972.

Barr, R., & Dreeben, R. How schools work: A study of reading instruction, in press.

Bennett, S. N. Teaching styles and pupil progress. Cambridge, Mass.: Harvard University Press, 1976.

Bidwell, C. E., & Kasarda, J. D. Conceptualizing and measuring the effects of school and schooling. American Journal of Education, 1980, 88, 401-430.

Bloom, B. S. Human characteristics and school learning. New York: McGraw-Hill, 1976.

Bloom, B. S. Implication of the IEA studies for curriculum and instruction. School Review, May 1974, 413-435.

Brown, W., & Saks, D. H. The production and distribution of cognitive skills within schools. Journal of Political Economy, 1975, 83, 571-593.

Brown, W., & Saks, D. H. Microeconomics of schooling. In D. C. Berliner (Ed.) Review of research in education, Vol. 9. Washington, D. C.: American Educational Research Association, 1980.

Burstein, L. The role of levels of analysis in the specification of educational effects. In R. Dreeben & J. A. Thomas (Eds.), Analysis of educational productivity, Vol. I: Issues in microanalysis. Cambridge, MA: Ballinger Press, 1980, pp. 119-190.

Burstein, L., Fischer, K., & Miller, M. D. The multilevel effects of background on science achievement at different levels of analysis: A cross-national comparison. Sociology of Education, 1980, 53(4), 215-255.

Burstein, L., Linn, R. L., & Capell, F. J. Analyzing multilevel data in the presence of heterogeneous within-class regressions. Journal of Educational Statistics, December 1978, 3(4) 347-383.

California State Department of Education. Evaluation Report of ECE, ESEA, Title I and EDY, 1974-75. Sacramento, Calif.: 1976.

Carroll, J. B. A model for school learning. Teachers College Record, 1963, 64, 723-733.

Comber, L. C., & Keeves, J. P. Science education in nineteen countries, International studies in evaluation (Vol. 1). Stockholm: Almqvist & Wiksell; and New York: Wiley, 1973.

Cronbach, L. J. Research on classrooms and schools: Formulation of questions, design, and analysis. Occasional Paper, Stanford Evaluation Consortium, Stanford, California, July 1976.

Davis, J. The campus as a frog pond: An application of the theory of relative deprivation to career decisions of college men. American Journal of Sociology, 1966, 72, 17-29.

Farkas, G. Specification, residuals, and contextual effects. Sociological Methods and Research, 1974, 2, 333-363.

Firebaugh, G. Groups as contexts and frog ponds: Some neglected considerations. Sociological Methods and Research, 1979, 7(4), 379-504.

Fisher, C. W., Filby, N. N., Marliave, R. S., Cahen, L. S., Dishaw, M. M., More, J. E., & Berline, D. C. Teacher behaviors, academic learning time and student achievement: Final Report of Phase III-B, Beginning Teacher Evaluation Study (Tech. Rep. VI). San Francisco: Far West Laboratory for Educational Research and Development, June 1978.

Greene, J. Choice behavior and its consequences for learning: An ATI study. American Educational Research Journal, 1980, 17(3), 291-302.

Harnischfeger, A., & Wiley, D. E. The teaching-learning process in elementary schools: A synoptic view. Curriculum Inquiry, 1976, 6, 5-71.

Hauser, R. M. Socioeconomic background and educational performance (Rose Monograph Series). Washington, D. C.: American Sociological Association, 1971.

Hauser, R. M. Contextual analysis revisited. Sociological Methods and Research, February 1974, 2(3), 365-375.

Husen, T. (Ed.) International study of achievement in mathematics. A comparison of twelve countries. New York: John Wiley & Sons, 1967.

Kelly, A. Girls and science (IEA Monograph Studies No. 9). Stockholm, Sweden: Almqvist & Wiksell International, 1978.

Klitgaard, R. E. Going beyond the mean in educational evaluation. Public Policy, 1975, 23, 59-79.

Lohnes, P. Statistical descriptors of school classes. American Educational Research Journal, 1972, 9, 547-556.

McDonald, F. P., & Elias, P. Final Report, Phase II (Beginning Teacher Evaluation Study). Princeton, N. J.: Educational Testing Service, 1976.

Murname, R. J. The impact of school resources on the learning of inner-city children. Cambridge, MA: Ballinger Publishing Co., 1975.

Rakow, E. A., Airasian, P. W., & Madaus, G. F. Assessing school and program effectiveness: Estimating teacher level effects. Journal of Educational Measurement, 1978, 15, 15-22.

Rosenshine, B. V. Formal and informal teaching styles: A review of S. N. Bennett, Teaching styles and pupil progress. American Educational Research Journal, 1978, 15, 163-169.

Schwille, J. Framework for IEA longitudinal study: Reader's guide to the draft questionnaires. Unpublished manuscript, Michigan State University, East Lansing, 1978.

Snow, R. E. Learning and individual differences. Review of Research in Education, 1977, 50, 273-289.

Sorenson, A., & Hallinan, M. T. A reconceptualization of school effects. Sociology of Education, 1977, 50, 273-289.

Stebbins, L. B., St. Pierre, R. G., Proper, E. C., Anderson, R. B., & Cerva, T. R. Education as experimentation: A planned variation model (Vol. IV-A, An evaluation of Project Follow Through). Boston: ABT Associates, April 1977.

Travers, K. J. Problem and Prospects in Cross-National Research: The Second IEA Mathematics Study as a Case Study. Presented at the annual meeting of the American Educational Research Association, Boston, MA, 1980.

Webb, N. Group Process: The key to learning in groups. In K. H. Roberts & L. Burstein (Eds), Issues in aggregation, Vol. 6, New directions for methodology of social and behavioral science. San Francisco, CA: Jossey-Bass, 1980, 77-88.

Wiley, D. E. Design and analysis of evaluation studies. In M. C. Wittrock & D. E. Wiley (Eds.), The evaluation of instruction: Issues and problems. New York: Holt, Rinehart, & Winston, 1970.

Wiley, D. E., & Bock, R. D. Quasi-experimentation in educational settings: Comment. The School Review, 1967, 75, 353-366.