DOCUMENT RESUME

ED 263 116 TM 850 006

AUTHOR	Burstein, Leigh
TITLE	Information Use in Local School Improvement: A Multilevel Perspective.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Apr 84
GRANT	NIE-G-84-0112-P1
NOTE	42p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984). In: Comprehensive Information Systems for Local School Improvement: A Reality-Test in Secondary Schools (TM 850 001).
PUB TYPE	Speeches/Conference Papers (150) -- Information Analyses (070)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Data Analysis; Educational Environment; Educational Research; *Evaluation Methods; High Schools; *Information Systems; Institutional Characteristics; *Instructional Improvement; *Research Methodology; *Systems Approach; Test Results
IDENTIFIERS	Multilevel Analytic Methods; *Systemic Evaluation; Unit of Analysis Problems

ABSTRACT

This is one of a series of reports based on an ongoing reality test of systemic evaluation for instructional decision making. This feasibility study is being carried out by the Center for the Study of Evaluation with the Laboratory in School and Community Relations at a suburban Los Angeles high school (called Site A). Viewing a school as a cultural/ecological system, systemic evaluation is a set of principles to guide local development of processes to gather, organize, and utilize information relevant to the needs and values of the system and its members. This paper examines methodological issues derived from a multilevel analytic perspective on local school improvement. A multilevel analytic methodology examines data from the different levels (student, classroom/teacher, school, district) of the educational system and their interconnections. It employs test analysis procedures that reflect instructionally sensitive performance variation. It links the conceptualization of an educational process to its measurement and analysis at various levels. It uses analytical procedures that potentially identify effects at and within each level of the educational system. Numerous tables and a three-page list of references conclude the paper.

Information Use in Local School Improvement: A Multilevel Perspective*

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

Paper prepared for the annual meeting of the American Educational
Research Association, New Orleans, LA April 27, 1984

* This report represents work done at the Center for the Study of Evaluation (Pursuant to a grant from the National Institute of Education). However, the opinions expressend herein are the responsibility of the authored do not necessarility reflect the position or policy of NIE and no official endorsement should be inferred.

School systems have had rich information environments for quite some time. One need only browse the list of information typically available in local school districts (see Table 1) to realize how much data is routinely gathered and maintained in some fashion. Until recently, however, many school districts lacked the financial and human resources (and often, the incentives) necessary to make the available information serve as a viable component of decision-making in ongoing, enduring school improvement efforts.

Obviously, times have changed as is evidenced in other papers in this symposium and related work (e.g., Bank and Williams, 1983). The collection, analysis and maintainence of both achievement and non-achievement data within a comprehensive information system in order to (1) examine the functioning and impact of existing school programs, (2) monitor key school "health" indicators and (3) plan, guide and examine new instructional improvement initiatives is becoming a common feature of the local educational scene. These activities are both a sign of the times (once again, education and educational improvement matter) and an indirect by-product of the technology explosion and the improved sophistication of LEA research and evaluation personnel that resulted from evaluation and testing requirements associated with compensatory education programs. If Lyons, Doscher, McGranahan & Williams (1978) were to replicate today their study of evaluation practices in school districts, lack of computer resources and expertise and staff technical skills would be much less severe problems. To a great extent the knowhow and wherewithal exist within school districts to make their available information useful for decision-making with respect to a variety of local issues and problems.

Now that the expertise and technology are possible, it seems appropriate to devote more attention to fine-tuning local school data analysis and reporting to make practices in these areas (a) more sensitive to the substantive decisions school personnel must make and (b) better reflect the methodological state of the art. To this end, this paper examines selected analytical issues that arise in making information obtained from the multiple levels (pupil/parent, teacher/class, school, district, community) of local school settings useful for decisions at the various levels.

Even after a local school community has decided to undertake school improvement and allocates the resources to develop information systems for use in their efforts, questions remain about how the relevant information should be treated analytically. Although most districts routinely collect much of the pertinent data for school improvement efforts, these data are seldom analyzed and reported in a manner consistent with extant knowledge about the possibilities and limits of information from multilevel social structures. Since the same data (e.g., from standardized achievement or competency tests; parent, student, teacher, and administrator surveys; archival and demographic records) can take on different meanings when analyzed and reported at different levels (e.g., class vs. school) and by different indicators at the same level (e.g., the school average vs. the proportion exceeding a specified level of mastery), comprehensive information systems need to be designed to make valid, pertinent knowledge accessible for constituents at the multiple levels. This concern for knowledge utility and accessibility may require that the same data be reconfigured (or simply reported differently) for different users. Moreover, multilevel analyses, albeit handled in a technically appropriate manner, must be reported in a form suitable for the analytically unsophisticated.

In the remainder of the paper, we briefly discuss several methodological issues derived from a multilevel analytic perspective on local school improvement. This discussion draws heavily from both my previous work on analysis of multilevel data (Burstein, 1980, 1981 in press; Burstein and Linn, 1982; Burstein, Linn & Capell, 1978; Burstein, Miller, & Linn, 1982) and from a longer conceptual synthesis that attempted to apply the general multilevel method framework to the local school improvement context (Burstein, 1983). The examples cited are taken from both applied research and from school district information and practices (both hypothetical and real).

The Nature of a Multilevel Perspective

Local school districts engaged in school improvement would appear to be settings particularly amenable to adopting a multilevel perspective regarding the collection, analysis, interpretation and reporting of information on school contexts, programs, and outcomes. To be internally consistent with this perspective, it is necessary to employ an analytical methodology that examines data from the different levels (student, classroom/teacher, school, district) of the educational system and their interconnections. Such a methodology incorporates both a multilevel conception and an accompanying willingness to disentangle effects from a variety of sources and levels (Figures 1 and 2 depict the data sources, data domains, and aggregation levels likely to be of interest in most school settings.) One must begin with a belief that no level of the educational system is uniquely responsible for the delivery of and response to schooling and thus substantive questions should rarely be confined to a single level (Burstein, 1980, in press; Rogosa, 1978). Thus a multilevel perspective and associated investigation focuses on the interface of individuals and the "groups" to which they belong and on the implications of this interface for understanding schooling.

There have been a number of syntheses of relevant research that focus on the theoretical, conceptual, and empirical bases for the impact of the multilevel character of educational systems on the measurement and identification of the antecedents and correlates of educational performance (Madaus, Airasian & Kellaghan, 1980; Barr & Dreeben, 1983; Bidwell & Kasarda, 1980a, 1980b; Burstein, 1980a, 1980b; 1983, in press; Cooley, Bond & Mao, 1981; Cronbach, 1976; Miller, 1981). These authors build a case that schooling can be better understood by, among other things,

- utilizing an array of group-level (class, school, etc.) indicators that are potentially sensitive to differential performance associated with differential resource allocation strategies
- employing test analysis procedures that are likely to reflect instructionally sensitive variation in performance
- linking a conceptualization of an educational process to its measurement and analysis at various levels
- using analytical procedures that potentially identify effects at and within each level of the educational system

Each of these points identifies analytical issues addressable from a multilevel perspective that becomes especially salient in information-rich school improvement contexts. In the remainder of the paper, we discuss and illustrate each point as it might arise as part of a school improvement effort. In doing so, we focus on group/social/organizational rather than individual/clinical uses of information for class-level, school-level and multi-school decision-making[1].

Alternative Indicators of Group Performance

Much of the analysis and reporting of achievement and non-achievement (e.g., attendance, taking of advanced course-work) indicators of educational performance is conducted at the group level (typically school or classroom). All too often these group-level analyses employ only measures of central tendency (such as means and medians, average percentages). But when one's purpose is to understand schooling and depict its consequences, such measures of central tendency can hide important differences in the distribution of pupil performance and educational experiences. Under many circumstances, the distribution of performance from an instructional setting is likely to be as informative about the operation of educational processes as the group's typical performance (Brown & Saks, 1975; Burstein, 1980; Burstein & Linn, 1982; Cooley & Lohnes, 1976; Klitgaard, 1975; Linn & Burstein, 1977; Lohnes, 1972; Spencer, 1983; Wiley, 1970).

The point here is that although a concern for achievement may drive instructional improvement efforts, it is important to keep in the local school context and be more refined in its objectives. Under usual schooling conditions, a focus on raising the performance level of students around the middle of the overall performance distribution (e.g., say 40th - 60th percentile) will yield the highest gains in mean performance[2]. Yet such a thrust ignores just the segment of students who have the greatest needs. A focus on the performance of the lowest quartile, on the other hand, devotes instructional resources in a manner likely to reduce the spread of performance (by establishing a performance floor or boosting more students over the minimal mastery point). Thus, the multilevel principle that group means do not account for all relevant group-level information should lead to context-sensitive analyses intended to monitor instructional improvement[3].

Examining Spread

The hypothetical data in Table 2 illustrates the value of multiple group-level indicators in simple comparisons of class-level performance. While the students in both classes started at the same level on the pretest and had the same mean performance on the posttest, the variation in posttest performance is much larger for Class 2. Instruction in this class led to differential increments in learning gains (some students learned more than their counterparts at the same pretest performance in Class 1 while others learned less) while the posttest performance patterns in Class 1 reflect a more uniform distribution of learning gains. One needs to know more about the specific circumstances to judge whether one class-level performance profile is to be preferred over the other, but the two profiles are definitely different and should be treated accordingly.

Comparing Distributions

The data reported in Table 3 (taken from Spencer, 1983) illustrate the importance of trying to capture the entire distribution of performance in comparing groups (schools in this case)[4]. If judgments about a school's effectiveness were based on a criterion score of 70, School C would be top ranked (because 20 percent of its students exceeded this score) while School A would be ranked at the top if the cutoff were either 45 or 40. This example might seem far-fetched, but it depicts what can happen when schools are judged by either the number of students taking advanced placement courses or their average SAT score versus judging schools according to the proportion of students passing a state's minimum competency or high school proficiency test.

Measuring School Effectiveness by Subgroup Comparisons

Another actual example derives from the recent school effectiveness literature. In his studies, Edmonds (1982) focussed on within-school differences between lower SES and higher SES students in the proportion of students achieving mastery of designated educational objectives. Thus, if a substantial proportion of a designated group within a school performed at the prescribed level or if the differences in proportions achieving mastery were approximately equal, the school was judged "effective". In essence, Edmonds' interest in the antecedents of effective achievement of low-income students caused him to depart from typical practice in school effectiveness studies of concentrating on school mean levels across all pupils.

Characterizing Survey Responses

The last example is taken from the reporting of responses to survey items collected as part of A Study of Schooling (Goodlad, Sirotnik, & Overman, 1979) and later provided as part of a feedback package for a school (Table 4). Considering the student responses in Exhibit A first, the mean response to every item provides little indication of the diversity of feeling expressed by students. A significant number of students strongly dislike each activity, while in the case of "working with the whole class", an equal number strongly like this mode of learning. Students indicated definite preferences that, if reflecting true feelings, dictate against concluding that either it doesn't matter or students are undecided.

A somewhat different point is illustrated by the teacher responses to the organizational problem solving and principal leadership dimensions. The dimensions yielded equal means, but teacher opinions are certainly more divided with regard to their beliefs about principal leadership. The staff cohesiveness dimension exhibits both a higher mean and the virtual absence of "disagreement" responses (only 3% scoring either 1 or 2). The means alone simply cannot capture what these data tell us about teacher perceptions.

There are other studies and other group-level indicators (e.g., Burstein & Linn, 1981; Burstein, Linn, & Capell, 1978; Burstein, Miller, & Linn, 1981) that offer possibilities for improving explanations of the relationships among educational contexts, processes and outcomes. How to adapt at least the logic, and perhaps the methods, of alternative group-level indicators for use in local school improvement activities clearly warrants further consideration.

Instructionally Sensitive Test Analysis

Clearly, the emphasis in test construction, analysis, and interpretation is on individual differences in both classic and IRT psychometric treatments of test data. Yet it would seem that a concern for the sensitivity of test performance to instructional experiences would require test analyses approaches that reflect the organization of instruction and the circumstances in which students receive instruction. There is ample evidence of the substantial variability across classrooms and schools in their content selection, emphasis, coverage, and method of coverage, not to mention the quality of instruction. Under these circumstances, multilevel examinations of the patterns of test performance (at the class and school level in addition to the student level) can be potentially valuable for detecting effects to background differences (e.g., prior learning, socieconomic and demographic differences), instructional coverage and emphasis, and instructional organization (e.g., grouping and pacing effects). When these separate effects can be identified, it may then be possible to construct measures and indices which are sensitive to the context factors of instruction and describe performance accordingly.

We illustrate multilevel apprachs to understanding and describing test performance in two ways. First we present a hypothetical example of the hazards of basing interpretations of group (school, class) test performance on horizontal aggregates of many subskills and competencies. This example is followed by a brief report from an investigation of class-level patterns of responses to test items that, in our view, illustrates what can be gained from a multilevel analysis of patterns of responses to test items.

Aggregating Over Test Content

The proper handling and interpretation of scores aggregated over individuals has received most of the attention in multilevel methods literature. Yet, in several respects, the logic holds as well for the content of instruction and of outcome measures. The choice between an emphasis on basic skills or on a broader array of knowledge has much in common with the decision about which group or level is of interest. Just as a focus on low income students at the school level dictates interest in certain indicators of performance (and perhaps disinterest in others), judgments of the success of school improvement efforts can depend on the chosen level of aggregation over the content of instruction. It can also depend on the form of measurement of the content.

The concept of level of aggregation and measurement of instructional content can be depicted as follows. Following a literal interpretation of the dictums of Title I, School A institutes a strong back-to-basics effort, adding more drill and practice activities (spelling quizzes, timed math drills, memorization and recitation of poems). The additional time School A devotes to these activities is obtained by foregoing most social studies, arts, science, and music instruction. School B, on the other hand, increases enrichment activities and attempts to enhance the breadth of its curriculum through dramatic play in its social studies work (e.g., various classes enact political campaigns prior to major elections, operate mock city governments, "live" through the experience of the Pilgrims, etc.). The teachers in School B tie in most lessons in reading, mathematics, and writing with these dramatic assignments but leave little time during school hours for drills on math, spelling and language arts facts.

Schools A and B make clear their instructional preferences. However, unless the array of non-teacher made tests (annual standardized tests, state assessment, district continuum) given in the schools are unusual, aggregate scores from these tests will differentially reflect the instructional emphasis and quality of the two schools. If two-thirds of a test's reading questions were devoted to word identification, vocabulary and spelling and its math questions to basic facts and computations, total test scores might make it appear that School A is more effective because its students uniformly mastered their narrower curriculum material that predominated in total scores, while School B's students performed more inconsistently on their facts and mechanics. (Presumably, School B would do much better on more complex comprehension, reasoning and interpretation material that is covered by a more limited portion of the test.)

Unfortunately, this fictional scenario probably occurs all too often in current efforts to determine the content of instruction and its measurement in school improvement efforts. The premise here is that the level of refinement in distinctions about instructional content should be an explicit choice and the measurement of the consequences of instruction should be sufficiently refined to reflect both desired and unintended content distinctions. Aggregation over content in the scores from multidimensional tests is potentially hazardous if the intent is to determine what has been taught and how well.

Examining Response Patterns at the Group Level

Examinations of the patterns of students' responses to test items across occasions and across groups (classes, schools, etc.) can be a particularly informative means of deriving explanations of test performance. A reanalysis of selected test item data from the Beginning Teacher Evaluation Study (BETS; Fisher et al., 1978) suggests how response patterns might provide information about instructional differences. Miller (1981, 1984) examined answers chosen by 123 students from 21 fifth-grade classrooms to the 15 items from the fractions subtest on two occasions (prior to (Occasion B) and following (Occasion C) most instruction in this subject area). He classified the test items into four subtopics: adding of fractions, subtracting of fractions, equating fractions, and solving fractions with algebraic unknowns (e.g., X/3 = 6/9. What is X?). Tables 5 through 7 present a subset of Miller's results.

First, the intercorrelations of item performances between classes and among students within the same class (Table 5) clearly indicate the effects of differential topic coverage across classes and within classes. Virtually all of the high item intercorrelations occur in the between-class analysis. This reflects differential coverage of topics across classes. Those classes which taught addition of fractions also taught subtraction and those classes with high (low) performance on addition also tended to exhibit high (low) performance on subtraction. The classes which taught algebraic unknowns apparently didn't teach (or unsuccessfully taught) addition and subtraction of fractions. On the other hand, there was virtually no intercorrelation among item performance within classes. There was very little tendency for a student to perform well across all items. Different students answered the various questions correctly[5].

Table 6 present results from a selected set of classes with distinctive response patterns. Since time allocated to fractions instruction per week is also reported, it is possible to also speculate as to whether classes received any fractions instruction at all. Note especially the contrasts in performance. For example, classes 8 and 5 had the same posttest score and approximately the same gain, but class 18 did so by covering and mastering every topic except addition of fractions (these students exhibited the same performance pattern at pretest so it is likely that the teacher simply did not check to determine whether students had mastered the mechanics of fraction addition after the pretest.) Class 5 likely depicts the more typical pattern of mastery of simple addition and subtraction of fractions and virtually no other content coverage.

Classes 3 and 16 represent another interesting contrast. These two lowest scoring classes got there in different ways. The teacher in Class 16 devoted a substantial amount of time to the coverage of fractions but virtually all the time seems to have been spent learning about algebraic unknowns. Class 3 devoted almost no time to the fractions topic (the majority of the class did not attempt to answer the questions at either pretest or posttest.) But there was some differentiated content coverage as a few students in Class 3 mastered all four subtopics.

Table 7 presents information derived from examining the class-level variation in the actual response alternative selected. This analysis clearly suggested systematic differences across classes in the types of errors students made. The two addition-of-fraction items demonstrate that students in several classes apparently never learned not to simply add numerators and denominators. Several classes did not appear to teach the expansion of fractions or taught fraction reduction in such a way that the students did not grasp its flip-side. The algebraic unknowns items exhibit the greatest variety of class-specific confusion. Obviously, some aspect of instruction is responsible for the systematic misunderstanding of how to do this type of problem. It may simply mean that in the absence of instruction on algebraic unknowns, students from a given class facing a novel task (for them) try to apply some other algorithm they have learned.

This example of investigating patterns of test performance barely scratches the surface of modern psychometric work on item response patterns (See, e.g., Harnisch, 1983; Harnisch & Linn, 1981; Sato, 1975, 1980; Sato & Kuto, 1979; Tatsuoka & Linn, 1983), much less recent advances in information processing models for test design and interpretation (e.g., Baker & Herman, 1983; Brown & Burton, 1978;

Birenbaum & Tatsuoka, 1982; Curtis & Glaser, 1983; Davis, 1979; Tatsuoka, 1983). Yet it is clear that the combination of this modern orientation toward test design, better psychometric indices of response patterns, and analyses that take into consideration the levels at which instruction is delivered can provide better information about test performance for decision-makers.

Multilevel Measurement of Educational Processes and Contexts

The principle that the same observable variable can measure different constructs at different levels of analysis is well-established (Burstein, 1980a, 1980b; Burstein, Fischer, and Miller, 1980; Capell, 1981; Cronbach, 1976; Sirotnik, 1979). A few examples serve to emphasize its ubiquity in educational research. Take, for instance, the standard measures of socioeconomic background typically found in studying schools. At the individual level, they may properly convey the parental investment in the individual child's learning. Once aggregated to the school level, social background measures also reflect the community context (e.g., wealth, urbanism, commitment to quality education) which in many cases conditions the resource allocations to schools. Within an educational level, relative social background positions students within a potential status hierarchy (e.g., a big fish in a small pond) That can affect their experiences (Burstein, 1980a, 1980b; Burstein et al., 1980). All three measures of social background may be important in understanding the experiences and performances of students but they do represent distinctly different mechanisms.

In a reanalysis of data from an observational study of the factors influencing student learning, Burstein (1980) demonstrated how the interpretation of a measure of the relative amounts of student learning tasks judged easy changed as the analysis shifted from the student to the class level. Students' success rates in learning tasks at the individual level captured proximal student ability and thus werepositively related to student performance. At the class level, this same observational variable reflected teachers' policies with regard to task difficulty and in many instances exhibited negative relationships with student outcomes.

The problems of change in variable meaning across levels are particularly evident in the literature on organizational and educational climate (e.g., Capell, 1979; Sirotnik, 1979). The distinction between a specific student's perception of classroom climate, which reflects both absolute and comparative aspects of individual personality and perception, and the average perception of the class, a normative measure of the instructional environment, is an important one. Whether the "organizational" or the "psychological" aspect of the climate is most salient in a given context is unclear. Capell (1981), for instance, construed aggregate responses of teachers within schools on scales purported to measure the degree of innovation and teacher influence as indicators of the atmosphere and organizational structure of the school program (See Table 8). In contrast, the individual teacher responses, relative to the responses of other teachers in the school, were interpreted as indicators of the teachers' sense of personal efficacy. That the effects of aggregated and individual measures on pupil outcomes were opposite in sign and consonant with expectations reinforces the need for a better understanding of how aggregation affects the measurement of program and process characteristics.

The studies cited above are important for our present purposes in two respects. First, they demonstrate the value of multilevel methods in educational research and evaluation. Second, and more importantly, the measures used in these studies -- socioeconomic background indicators, survey responses from students, teachers and parents, classroom observation data -- represent typical information about educational processes and contexts that are or can be gathered in local educational settings. Apparently, these measures can serve as indicators of a variety of constructs -- home resources, community resources, organizational structure and atmosphere, personal efficacy, classroom and school climate, appropriateness of content, student level of functioning -- that are important in understanding and improving schooling when the linkage between the level of aggregation of measures and the construct of interest is clear.

The relevance of these concerns about the shift in variable meaning across levels is particularly pertinent to a variety of investigations of the effects of schooling that a number of local school districts have undertaken. Several school districts have conducted their own studies of school effects (e.g., Kean, Summers, Raivetz & Farber, 1979 (Philadelphia); Ramey, Hillman & Mathews, 1982 (Seattle)) and school effectiveness (e.g., White & Kemp, 1976 (Atlanta); Gastright, 1977 (Cincinnati)). The methodology they employ mirrors the practices of large-scale investigations of school effects and school effectiveness. The LEA-based studies seem no more nor less resistant to problems in measuring the variables of interest at the appropriate levels, and incorporating them properly in their analyses, than their large-scale, multi-site counterparts (see Burstein, 1980; Madaus, Airasian & Kellaghan, 1980; and Purkey & Smith, 1982 for discussions of methodological problems with these types of studies). If local educational agencies use these investigations to guide their decisions about instructional improvement programs and other school renewal activities, then one would hope that inattention to specific concerns about appropriate level of measurement and its relevance to construct-indicator match would have limited impact.

Analytical Methods for Disentangling Multilevel Effects

Two recent classroom studies of beginning reading demonstrate the value of decomposing the variation of students' instructional experiences and performance into variation associated with subgroups within classrooms. Barr & Dreeben (1983) focussed on content coverage as a variable and found that it varied mainly between reading groups and not between teachers. Not surprisingly, students performance also varied primarily between reading groups.

In the Texas First Grade Reading Group Study (Anderson, Evertson & Brophy, 1979), the focus was on teaching behavior variables such as teachers' selection of students to read (e.g., non-volunteer selections), types of student responses to teacher questions, and types of teacher feedback. In a secondary analysis of the study data, Martin, Anderson, and Veldman (1980) decomposed the variation in student achievement into effects of teacher behaviors at three levels: students within reading groups, reading groups within classes, and classes. Most of the significant relationships were for students-within-reading groups. Also teachers tended to change their selection strategies across reading groups within their class. An analysis of class means only would have missed the effects of teachers' differential activities across reading groups and the differential impact of teacher behaviors on the members of specific reading groups.

There is no need to provide additional rationale for and describe developments in analytical methods for disentangling multilevel effects. As with the studies of reading, the substantive investigations described in earlier sections typically combined a better conceptualization of the multilevel character of educational data with analytical machinery adapted to the substantive questions of interest rather than molding the theory to meet the conditions of the statistical procedures. Generally, the analytical procedures employed were the familiar ones, but these tools were used in a variety of ways that better mirrored the process of schooling. Typically, better analysis of multilevel educational involve disentangling influences at and within each level of the educational system by conducting multiple analyses or a common analysis with measures collected from (or aggregated to) multiple levels.

Currently, methodological research on methods for analyzing multilevel data is focussing on relatively sophisticated procedures that require stronger assumptions but at the same time, are more robust to typical shortcomings in social science data (e.g., missing data, asymmetrical distributions, measurement error, heteroscedasticity, resistance to outliers. Relevant work is reported in Aitken, Anderson, and Hinde, 1982; Burstein & Gustafsson, in progress; Goldstein, In progress; Mason, Wong & Entwistle, 1983; Rachman & Wolfe, 1983; Schneider & Treiber, 1984. While little of this work will have direct bearing on routine analyses to guide local school improvement, they may provide a better means for re-interpreting results from studies of the effects of schooling in ways that are more consonant with the perceptions of local school personnel. If so, such reanalysis would provide better support for effective local practices and less ammunition for school critics.

Concluding Comments

The value of a multilevel perspective for understanding the effects of schooling is becoming a more commonly held perception across a wide array of educational professionals (researchers, policy makers, and practitioners). We have attempted to illustrate several ways in which such a perspective might lead to more sensitive and sensible data analysis that are better suited to school improvement efforts.

To some degree, our efforts bog down when we attempt to shift from studying and understanding the concept of multilevel analysis and turn to its practice. School cultures involve a variety of sources of information and a number of constituencies (e.g., teachers, counselors, administrators) with clearly demarcated responsibilities that might be able to use properly collected, appropriately analyzed and routinely accessible information. But little is known about how school building personnel operate in a context with high-quality, timely, pertinent information, either because these conditions do not exist or no one has yet documented how school-level personnel respond under such ideal information conditions. If the other symposium papers are indicative of what is now possible with respect to the use of comprehensive information systems in schools, we won't have to wait very long to determine whether the promise of a multilevel perspective toward the analysis and reporting of school data is real or illusory.

FOOTNOTES

[1] This distinction is a useful one. Individual/clinical uses of information include such activities using test data for individual diagnosis of learning problems, placement decisions, individual student counseling and guidance activities, administrator supervision of individual teachers and similar individual personnel matters (hiring, course asignments, promotions, etc.). Group/social/organizational uses refers to the myriad of ways in which data from individuals are aggregated and organized to characterize/depict/understand the functioning and behavior of groups of individuals. Class-level test performance and background profiles, subgroup (e.g., by ethnicity, sex, grade-level, curriculum track, SES) information; course enrollment and course-taking patterns across subject matters and subgroups of students are all examples of the latter. Most of the analytical developments pertinent to this paper deal with the latter type of use.

[2] There are both substantive and technical reasons for this. Substantively, students at higher percentile levels have a low ceiling; that is, much of the new material they might learn is not likely to be reflected in substantial improvements in the test performance because this material typically is not well-represented on the test. Low performance, on the other hand, may require substantial resources to boost performance above, say, the 50th percentile and thus draw off resources from students in other parts of the distribution. For example, concentration on the skills needed by low

performers may lead to more wait time and inefficiency for high performers (unless the latter are allowed to "work ahead on their own"). Performers in the middle of the distribution are likely to benefit from the focus because they have sufficient room to grow and may simply require a bit more targeted instruction to clear up certain misconceptions (see later example about fraction addition) or to acquaint them with topics not previously covered.

The technical side of the argument is the well-known relationship between raw score points and percentiles. It takes more raw score change to move up a given number of percentile points in the tails than in the middle of a normal distribution. Thus gains in knowledge in the middle of the distribution boost the average percentile more rapidly than gains in either the upper or lower tails.

[3] One does not have to be devoted to compensating for the inadequate performance of low income pupils to derive benefits from interest in the distribution of performance rather than simply its level. In their annual reports, the California Asessment Program provides schools with the quartile distributions of the performance of 3rd grade students along with a variety of mean indicators (overall and for various demographic subgroups). A former principal of a suburban, typically higher performing school pointed out that while the school's overall performance each year (typically above 70 per cent correct, which is above the 90th percentile statewide) didn't tell him anything, he did keep track of the number of children that fell in the lower quartile each year because this meant that there were still

students who needed to improve. Thus, even high achieving schools can benefit from an awareness of the functioning of their weakest students and school mean performance doesn't typically capture this type of information.

[4] Spencer's paper (1983) considers a number of statistical problems associated with the typical use of test scores for comparisons of outcome differences among groups. He highlights the problems associated with the ordinal properties of most metrics used to measure outcomes and presents a case for switching to indices of "stochastic ordering" when attempting such comparisons. This shift would certainly be in the direction of maintaining more distributional information in group-level analyses.

[5] The exceptions to the lack of correlation across items within classes are typically for those items that are essentially parallel (e.g., items 1 and 2; item 6 and 7; and items 11, and 13) or involve analogous straightforward topics such as addition and substraction items containing a common denominator.

Table 1

The Types of Information Routinely Collected (or collectable) in School Districts

A. Demographic/Archival

1. Student demographics--age, sex, ethnicity, home language, parental occupations and employers, eligibility for AFDC, reducted price lunches, medical histories, home address, mobility (how long in particular residence) parental education, family size
2. Teacher and building-level administrator backgrounds -- age, education, previous employment and educational history, special certification and subject-matter expertise
3. School building characteristics -- information about physical plant (e.g., age, capacity, particular resources),
4. Student body and community composition--ethnic composition, neighborhood wealth, community involvement in neighborhood schools (e.g., PTA membership)

B. Financial

5. Payroll expenditures
6. Materials and supplies
7. Equipment
8. Maintenance
9. Special programs (e.g. entitlement programs, staff development, remedial services, counseling and guidance)
10. Transportation
11. Safety and Security

C. Testing

12. Standardized norm-referenced tests
13. Criterion referenced testing
14. Minimum competency and proficiency testing
15. Group and individual ability and aptitude testing -- done typically to determine pupil eligibility for special programs and placement decisions
16. Teacher-made tests and curriculum embedded tests

D. Program Characteristics and Participation

17. Special program participation -- availability and staffing of special programs at local school sites
18. Curriculum information -- curricular packages and texts used in classrooms, topic coverage from continuum (assumed and measured)
19. Course taking patterns -- information from student cummulative records and from prescribed offerings
20. Grading practices -- teacher reports of student grades

E. Student Performance, Participation, and Behavior

21. Grades by content area
22. Participation in extracurricular activities by types
23. Awards -- e.g., scholarships
24. Absenteeism and tardiness
25. Reported disruptive and inappropriate behavior

F. Affective, Attudinal, and Observation Information

26. Student responses to surveys about class and school environments and other aspects of their educational experience
27. Teacher measures of classroom and school climate and activities
28. School building administrator measures of school climate and activities
29. Parental surveys of perceptions and support of school activities
30. Parental participation in school activities (e.g., volunteers, fundraising attendance at school functions, scheduled conferences)
31. Administrator observations and evaluations of teachers
32. Teacher observations of other teachers
33. District personnel's observation and interviews of building personnel
34. Surveys of graduates to determine occupational and educational status
35. Information about student dropouts

G. District Evaluation Reports

36. Routine annual reports to board and federal and state agencies

37. Evaluation of specific educational changes

38. Instances of local school assistance by type and disposition

[6] Source Burstein, L. The Use of Existing Data Bases in Program Evaluation and School Improvement (1983)

Table 2

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	$\bar{X}_{\cdot 1} = 3$	$\bar{X}_{\cdot 2} = 3$	$\bar{Y}_{\cdot 1} = 4$	$\bar{Y}_{\cdot 2} = 4$
Standard Deviations	1.08	1.08	1.08	2.17

Source: Burstein, L. & Linn R.L. Analysis of Educational Effects from a Multilevel Perspective:Disentangling Between- and Within-Class Relationships in Mathematics Performance, CSE Report No. 172, University of California Los Angeles, Center for the Study of Evaluation, 1982

Table 3 . Percent of Students Scoring at Least X

	Score x				
School	40	45	50	55	70
A	90	80	60	50	10
B	85	75	55	48	14
C	83	70	65	48	20
D	71	60	40	20	3
⋮	⋮	⋮	⋮	⋮	⋮

Source: Spencer B. D. On interpreting test scores as social indicators: Statistical considerations, Journal of Educational Measurement, 1983, 20(4), 317-334.

Table 4

Examples of Responses to Survey Items where Mean response is insufficient to reflect distribution of responses.[a]

Exhibit A— Responses of 26 students from a single secondary school class.

	Percentage of students responding...				
	Like very much	Like somewhat	Dislike somewhat	Dislike very much	Mean
Working with the whole class	23	38	15	23	2.4
Tell in my own words what I have learned.....	04	23	31	42	3.1
Do word problems........	04	54	19	23	2.6
Do research and write reports...............	04	23	31	42	3.1

Exhibit B— Responses of 34 teachers from a single school to items measuring teachers perceptions of the work environment. The 77 items were combined into 3 scales (dimensions of the work environment) which were labeled "organizational problem solving", "principal leadership" and "staff cohesiveness". Items were answered on a six-point agreement scale and the school mean and distribution of teacher scores (average response to items from a given dimension) are reported below:

Dimension	Mean	Teacher Distribution (%)					
		1	2	3	4	5	6
1. Oranizational Problem Solving	3.4	3	12	41	29	12	3
2. Principal Leadership	3.4	12	18	24	26	9	12
3. Staff Cohesiveness	3.7	0	3	44	35	15	3

[a] Data and questions taken from an example feedback package from A STudy of Schooling (Goodlad, Sirotnik et. al.) which also appeared as Appendix B in Sirotnik & Burstein (1983).

Table 5

Item intercorrelations between classes (lower triangle) and within classes (upper triangle) on occasion C.

	Subtraction					Addition					Equating			Algebraic Manipulation	
ITEM	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		.75	.07	.33	-.03	.51	.29	.09	.01	.23	.23	.21	.22	.26	.20
2	.86		.00	.29	-.08	.54	.33	.07	-.00	.16	.19	.12	.02	.29	.18
3	.19	.28		.30	.06	.17	.29	.26	.25	.23	.29	.03	.05	.16	.06
4	.62	.53	.50		.03	.25	.30	.23	.22	.19	.28	.05	.15	.40	.26
5	.43	.41	.36	.39		-.15	-.00	.37	.40	-.01	.19	.18	.19	.16	-.01
6	.68	.42	.34	.43	.27		.59	.14	.15	.36	.19	.08	.07	.05	.19
7	.52	.29	.55	.48	.38	.86		.29	.25	.50	.19	.16	.08	.16	.18
8	.49	.29	.30	.48	.42	.66	.64		.67	.24	.35	.29	.23	.35	.20
9	.53	.37	.33	.45	.69	.64	.59	.75		.19	.21	.24	.16	.23	.24
10	.56	.28	.05	.40	.49	.61	.62	.53	.69		.12	.06	.13	.04	.23
11	.55	.45	.33	.56	.64	.42	.45	.47	.58	.51		.50	.39	.27	.23
12	.41	.37	.60	.74	.52	.26	.38	.57	.41	.23	.69		.38	.29	.35
13	.50	.51	.19	.45	.60	.26	.16	.59	.51	.29	.56	.69		-.17	.29
14	.18	.07	-.07	.36	.31	-.10	-.12	.41	.35	.12	.47	.50	.55		.37
15	.29	.18	-.05	.33	.39	.10	.08	.48	.53	.33	.64	.48	.49	.81	

Source: M.D. Miller, Item Response and Instructional Coverage, 1984.

Table 6

Classrooms exhibiting distinctive class-level patterns of performance on fraction items.[a]

Class #	Allocated Time Mean[b]	Allocated Time Rank[c]	Posttest Mean[d]	Posttest Rank[e]	Gain Mean[f]	Gain Rank[g]	Nature of the item response pattern
18	119	1	7.5	10.5	5.0	7	Mastered equating, algebraic manipulation, and simple subtraction. Only problem at posttest was addition of fractions (adding numerator and denominator) where only 2 correct out of 30 responses (6 students responding to 5 items).
5	77	5	7.5	10.5	5.2	5	Students can do simple addition and subtraction items but not more complicated ones. No work on equating. Some coverage of algebraic manipulation but basically don't understand the concept.
27	114	2	12.5	1	6.5	1	No problems at posttest except on algebraic manipulation.
12	87	4	11.3	3	6.3	2	Most mastered everything but 1 or 2 didn't master any fraction topic (possible differential coverage).
11	77	6	12.3	2	1.7	15	Lots of time spent on fractions with not much new learning, at least on the topics measured by the test.
8	26	14	9.0	7	5.6	3	Success on topics covered (subtraction,equating, simple addition). Addition other than simple common denominator mislearned or not taught.
26	46	10	6.4	13	5.0	7	Fraction curriculum at low level. Most learned only addition and subtraction with common denominator. One or two students learned more.
14	29	12	5.3	15	3.7	9	Differentiated teaching and learning. Some mastered everything, some mastered addition and subtraction with common denominator, others mastered (probably covered) nothing.
16	52	7	2.8	21	1.5	16.5	Students in this class only mastered algebraic manipulation They did not answer any other questions on either pretest or posttest (with one or two exceptions).
3	3	19	4.8	20	3.3	10.5	Differentiated content coverage,almost all on algebraic manipulations.some students mastered most topics.

[a] Based on scores of 123 fifth graders from 21 classrooms in the Beginning Teacher Evaluation STudy.These analyses are reported in Miller (1984).

[b] minutes per week spent on fractions during period between protest and posttest.

[c] Based on ranking of mean time allocated to fractions across the 21 classrooms.

[d] Based on posttest average.for a sample of approximately 6 students from each class. Maximum possible score was 15.

[e] Based on ranking of mean achievement score at posttest for the 21 classes.

[f] class mean difference between total scores at the pretest and posttest.

[g] Based on ranking of the mean gains for all 21 classrooms.

BEST COPY AVAILABLE

Tible 7

Test Items exhibiting distinctive class-level patterns of student performance*

Question	Alternatives				Nature of the Response Pattern
$5\frac{5}{8} - 4\frac{1}{4}$ Pretest p-value=.11 Posttest p-value=.35	$1\frac{3}{8}$	$1\frac{4}{8}$	2	$1\frac{4}{12}$	As students learned about subtracting fractions, many did not learn how to obtain a common denominator before subtracting. This resulted in somè classes systematically choosing alternative 2, probably because they had been taught how to handle fractions with a common denominator (occurred in 5 classes at posttest). 3 classes exhibited mastery at posttest.
$\frac{2}{3} + \frac{1}{3} =$ Pre p-value =.11 Post p-value=.26	$\frac{3}{6}$	3	$\frac{2}{9}$	1	The most common error for both questions is adding both the numerator and denominator. This problem was not resolved at the posttest for several classes (9 classes on the first item, 13 on the second). Both questions require change to mixed fractions which students in many classes apparently were not taught.
$\frac{3}{7} + \frac{5}{7} =$ Prep-value = .07 Post p-value=.24	$\frac{28}{35}$	$\frac{8}{14}$	$1\frac{1}{7}$	$\frac{4}{7}$	Four classes exhibited mastery of the first item and 2 classes exhibited mastery of the second item at posttest
$\frac{1}{2} =$ Pre p-value = .32 Post p-value= .55	$\frac{2}{3}$	$\frac{3}{6}$	$\frac{2}{5}$	$\frac{4}{9}$	Students in some classes could not expand fractions at the posttest (5 classes on each item). They simply added 1 to both numerator and denominator.
$\frac{2}{3} =$ Pre p-value = .15 Post p-value= .32	$\frac{8}{12}$	$\frac{3}{4}$	$\frac{3}{9}$	$\frac{5}{6}$	8 classes mastered the first item at posttest and 2 classes exhibited mastery of the second item.
What does N equal? $\frac{2}{7} = \frac{N}{21}$ Pre p-value = .20 Post p-value= .33	3	16	7	6	Lots of class-specific confusion. Only 1 class exhibited mastery. Four classes systematically chose alternative 1, 2 classes systematically chose alternative 2 and 1 class systematically chose alternative 3 at the posttest. 9 classes mastered the somewhat easier item 3/8 = 6/N (posttest p-value =.53) with only limited systematic errors.

*These are test results from selected fraction items given to 123 students in 21 fifth grade classrooms before and after instruction on this topic(in most classes). The data were collected as part of the Beginning Teacher Evaluation Study (Fisher et. al., 1978). The results reported here are based on a dissertation by Miller (1981) and are reported more thoroughly in Miller (1984).

Table 8 REGRESSION OF CLASS LEVEL STUDENT READING ACHIEVEMENT ON TEACHER AND MINISCHOOL LEVEL SURVEY VARIABLES[a,b]

Variable Name	Multilevel Regression		Class Level Regression
	Teacher Level	Minischool Level	
Staff Cohesion	-.225 (.487)	.679 (.849)	.197 (.536)
Common Minischool Policies	-.336 (1.445)	.360 (1.155)	-.057 (.371)
Teacher Autonomy	1.652 (1.601)	-2.256 (1.323)	.563 (.696)
Teacher Influence	.660 (1.714)	-1.330 (2.292)	-.135 (.508)
Principal Influence[c]	.545 (1.518)	-.961 (1.843)	.004 (.000)

[a] Unstandardized regression coefficients, t statistics in parentheses.

[b] The regression equation was estimated using the method of weighted least squares:

$$\underset{\sim}{\beta} = \{\underset{\sim}{W}'(\underset{\sim}{X}'\underset{\sim}{X})\underset{\sim}{W}\}^{-1}\ \underset{\sim}{W}'\underset{\sim}{X}'\underset{\sim}{y}\underset{\sim}{W}\ ,$$

where

$$\underset{\sim}{W} = tr(\underset{\sim}{n}_i)^{-1}k\underset{\sim}{n}_i = N^{-1}k\underset{\sim}{n}_i\ ;$$

$\underset{\sim}{\beta}$ is the vector of regression coefficients; $\underset{\sim}{X}$ is the matrix of independent variables; k is the number of classes; N is the total number of students; and $\underset{\sim}{n}_i$ is a diagonal matrix of class sizes. Use of the matrix $\underset{\sim}{W}$ insures that each classroom in the analysis will be weighted by the number of students contained in it, while the overall degrees of freedom for classes will be preserved.

[c] This variable was measured such that a negative coefficient represents greater influence.

Source: Capell, F.J. A study of alternatives in American education, Volume VI: Student outcomes at Alum Rock 1974-1976, R-2170/6-NIE, Santa Monica, CA: Rand Corporation, July 1981.

Data Sources		Data Domains (Examples Only)			
		Personal	Class	School	Schooling
Data Sources	*Teachers	• Demography • Reasons for entering education profession • Teaching experience • Educational beliefs	• Relative amounts of time spent on instruction, behavior control, and routines • Use of behavioral objectives • Frequency of certain learning activities	• Relative importance of school functions (social, intellectual, personal, and vocational) • School "climate" or work environment • Major problems • Equality of education (ability, race, sex)	• Desegregation • Fiscal support of public education • Teachers unions • Minimum competency • Role of global education in the schools
	Students	• Demography • Self-concept • Educational aspirations	•Relative amounts of time spent on instruction, behavior control, and routines • Difficulty of class content • Frequency of certain learning activites • Class "climate"	• Relative importance of school functions • Evaluative rating • Major problems • Equality of education • Adequacy of counseling services • Subject-area preferences	• Desegregation • Role of job experience in schools • Value of schools
	Parents	• Demography • Years lived in community • Political beliefs		• Relative importance of school functions • Evaluative rating • Major problems • Equality of education • Involvement in activities and decision making • Objectionable learning materials	• Desegregation • Fiscal support of public education • Teachers unions • Teachers' salaries • Minimum competency • Role of global education in schools
	Classroom* (Teacher/Student Interaction)		• Relative amounts of time spent on instruction, behavior control, and routines • Use of corrective feedback • Use of open versus closed questions • Instructional time, spent with total class versus individual versus groups		

*Data were collected on this data source through observation. For the purposes of this conceptualization, observers are being treated not as a data source, but as part of the data collection *method*, just as questionnaire and/or interview methods were used in collecting data from teachers, students, and parents.

Figure 1

The Schooling Terrain: Map One

SOURCE: Goodlad, Sirotnik & Overman, 1979

DATA DOMAINS

AGGREGATION LEVELS		Personal (Individual)			Instructional (Classroom)			Institutional (School)			Societal (Schooling)		
	Data Categories:	C	A	M	C	A	M	C	A	M	C	A	M
	Data Sources:												
Individual	Students												
	Teachers												
	Administrators												
	Parents												
Class	Students												
	Teachers												
	Administrators												
	Parents												
	Classroom												
School	Students												
	Teachers												
	Administrators												
	Parents												
	Classrooms												
	School												
District	Students												
	Teachers												
	Administrators												
	Parents												
	Classrooms												
	Schools												
	District												

Data Categories:

C = Circumstances

A = Activities

M = Meanings

Figure 2

The Schooling Terrain: Map Two

Source:Sirotnik K. & Burstein L. Systemic Evaluation, Los Angeles: University of CAlifornia Los Angeles, Center for the Study of Evaluation, October 1983.

BEST COPY AVAILABLE

REFERENCES

Barr, R., & Dreeben, R. How schools work: A study of reading instruction. University of Chicago, 1983.

Bank, A., & Williams, R. C. Annual report - Evaluation design: Organizational study. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1981.

Berman P., & McLaughlin, M. W. Federal programs supporting educational change Vol III: Implementing and sustaining innovations. R-1589/8-HEW, Santa Monica, CA: Rand, 1978

Bidwell, C. W., & Kasarda, J. D. Problems of multilevel measurement: The case of school and schooling. In K. Roberts & L. Burstein (Eds.), Issues in aggregation, No. 6, New directions in methodology for social and behavioral research. San Francisco: Jossey-Bass, 1980a.

Brown, J. S., & Burton, R. R. Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 1978, 2, 155-192.

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

Burstein, L. Using multilevel methods for local school improvement: a beginning conceptual synthesis, Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1983.

Burstein, L. Units and levels of analysis, International Encyclopaedia of Education, London, England: Pergamon Press, in press.

Burstein, L. The Analysis of Multilevel Data in Educational Research and Evaluation, in E. Berliner (ed) Review of Research in Education, Volume 8, Washington, D. C.: American Educational Research Association, 1980, 158-233.

Burstein, L. Analysis of multilevel data in educational research and evaluation. In D. Berliner (Ed.), Review of Research in Education, Vol. 8, Washington, D.C.: American Educational Research Association, 1980c, 158-233.

Burstein, L. The use of existing data bases in program evaluation and school improvement. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1983.

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. 1: Issues in microanalysis. Cambridge, MA: Ballinger, 1980b, 119-190.

Burstein, L., Fischer, K., & Miller, M. D. The mutilevel 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. Analysis of educational effects from multilevel perspective,: Disentangling between-and within-class relationships in mathematics performance, CSE Report No. 172. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1982.

Capell, F. J. A study of alternatives in American education, Volume VI: Student Outcomes at Alum Rock 1974-1976, R-2170/6-NIE, Santa Monica, CA: Rand Corporation, July 1981.

Cooley, W. W., & Lohnes, P. R. Evaluation research in education: Theory principles and practices. New York: Irvington Publishers, Inc., 1976.

Cooley, W. W., Bond, L. & Mao B-J, Analyzing multilevel data in R. A. Berk (ed.), Educational evaluation methodology: The state of the art, Baltimore, MD.: The Johns Hopkins University Press, 1981, 32-63.

Cronbach, L. J. Research in classrooms and schools: Formulation of questions, design, and analysis, with the assistance of J. E. Deken & N. M. Webb, Occasional Paper, Stanford Evaluation Consortium, 1976.

Edmonds, R. R. Programs of school improvement: An overview. Paper presented at the National Invitational Conference on Research on Teaching: Implicatons for practice, Warrenton, VA, February 1982.

Gastright, J. Some empirical evidence of the comparability of school unit residuals based on achievement and non-achievement variables. Paper presented at annual meeting of American Educational Research Association, New York, 1977.

Harnisch, D. L. Item response patterns: Applications for educational practice. Journal of Educational Measurement, 1983, 20, 191-206.

Harnisch, D. L. & Linn, R. L. Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 1981, 18(3), 33-46.

Kean, M. H., Summer, A. A., Raivetz, M. J., & Farver, I. J. What works in reading? Philadelphia, PA: School District of Philadelphia, 1979.

Klitgaard, R. W. 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, 574-556.

Lyon, C.D., Doscher, L., McGranahan, P., & Williams, R. Evaluation and school districts. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, CA, December 1978.

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

Martin, J., Veldman, D. J., & Anderson, L. M. Within-class relationships between student achievement and teacher behaviors. American Educational Research Journal, 1980, 17(4) 479-490.

Miller, M. D. Between-group differences in item response patterns. Unpublished doctoral dissertation, University of California, Los Angeles, 1981.

Miller, M. D. Measuring between-group differences in instruction. Unpublished doctoral dissertation, University of California, Los Angeles, 1981.

Miller, M. D. Item response and instructional coverage, unpublished paper, 1984.

Purkey, S. C., & Smith, M. S. Effective schools: A review. Paper presented at National Invitational Conference on Translating Research on Teaching into Practice, Arlington, VA, 1982.

Ramey, M., Hillman, L., & Mathews, T. School characteristics associated with instructional effectiveness. Paper presented at the Annual Meeting of the American Educational Research Association, March 1982.

Sato, T., & Kurata, M. Basic S-P score table characteristics, NEC Research & Development, 1977, 47, 64-71.

Sato, T. S-P table analysis--analysis and intepretation of test scores. Meiji-Tosho Publishing Co., Tokyo, 1975 (in Japanese).

Sato, T. The S-P Chart and the caution index. Tokyo: Nippon Electric Company, 1980.

Sirotnik, K. A. Psychometric implications of the unit-of-analysis "problem" (with examples from the measurement of organizational climate). Journal of Educational Measurement, 1980, 17, 245-281.

Sirotnik K. A., Burstein, L. & Thomas, C. Systemic evaluation, Los Angeles: Center for the Study of Evaluation, University of California Los Angeles, October, 1983.

Spencer B. D. On interpreting test scores as social indicators: statistical considerations, Journal of Educational Measurement, 1983, 20(4), 317-334.

Tatsuoka, K. K. & Linn, R. L. Indices for detecting unusual patterns: Links between two general approaches and potential applications. Applied Psychologocal Measurement.

White, B. F., & Kemp, D. M. Performance information: Did it make a difference? The Atlanta experience. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

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

Wisenbaker, J., & Schmidt, W. The structural analysis of hierarchical data. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.