DOCUMENT RESUME

ED 212 649 TM 820 004

AUTHOR	Miller, M. David; Burstein, Leigh
TITLE	Evaluation Design Project: Multilevel Interpretation of Evaluation Data Study.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (ED), Washington, D.C.
PUB DATE	Nov.81
GRANT	NIE-G-80-0112
NOTE	299p.; For related document, see ED 211 584; Appendices A, B, and D are marginally legible.
EDRS PRICE	MF01/PC12 Plus Postage.
DESCRIPTORS	Error Patterns; *Evaluation Methods; Item Analysis; Models; *Outcomes of Education; Program Effectiveness; Quasiexperimental Design; *Statistical Analysis; *Test Construction; Testing Problems; Test Validity
IDENTIFIERS	Beginning Teacher Evaluation Study; *Instructional Sensitivity; *Unit of Analysis Problems

ABSTRACT

Two studies are presented in this report. The first is titled "Empirical Studies of Multilevel Approaches to Test Development and Interpretation: Measuring Between-Group Differences in Instruction." Because of a belief that schooling does affect student achievement, researchers have questioned the empirical and measurement techniques used to evaluate the effects of schooling on student achievement. One possible shortcoming of the major standardized norm-referenced achievement tests is their failure to take into account the nesting of units in the educational system. This study uses item data at the class level to construct tests. It was found that selecting items from an index of discrimination between groups did lead to scales more sensitive to instructional differences. From the analysis of patterns of item response, it was found that not only did patterns of correct and incorrect item response vary as a function of class membership, but that the patterns of response reflect substantively meaningful differences in instruction. The second study, "State of the Art Methodology for the Design and Analysis of Future Large Scale Evaluations: A Selective Examination," reviews how recent methodological advances might be incorporated in future large-scale evaluations. Specifically, structural equation modeling and selection modeling and related issues in analysis of quasi-experimental data are examined. (Author/BW)

Center for the Study of Evaluation

UCLA Graduate School of Education
Los Angeles, California 90024

Deliverable - November 1981

EVALUATION DESIGN PROJECT: MULTILEVEL INTERPRETATION OF EVALUATION DATA STUDY

Annual Report

Leigh Burstein, Study Director

Grant Number

NIE-G-80--0112

P5

CENTER FOR THE STUDY OF EVALUATION
GRADUATE SCHOOL OF EDUCATION
University of California, Los Angeles

During the period of December 1, 1980 - November 30, 1981 two areas of inquiry were underaken by staff of the Multilevel Interpretation of Evaluation Data Study. Most of the study effort was devoted to an in-depth empirical study of multilevel approaches for test development and interpretation. A report detailing the results of this investiation is attached (Appendix A) and a summary of its purpose, methods, results, and implications is provided below.

The remaining work effort was devoted to an examination of how recent methodological advances might be incorporated in future large-scale evaluations. The actual activities varied somewhat from original plans. It is possible at this point to describe the major analytical developments and to consider their possible utility in large-scale program evaluations in education. A report detailing current work is attached (Appendix B) and a brief summary is provided below. At present there are no plans to continue this line of inquiry as a separate study when the MIED study activities are transferred to the CSE Methodology Program. We will continue to monitor methodological developments potentially relevant to large-scale program evaluation, but will not make this a primary thrust of our work given the limited resources available.

Empirical Studies of Multilevel Approaches to Test Development and Interpretation

Review and Rationale

During the past several years, CSE personnel have been working on the applicability of multilevel methods to test development and interpretation. An initial report (Miller & Burstein, 1979) detailing conceptual models for applying multilevel analysis principles to test development and interpretation was submitted in November 1979. However, it was clear that we had only begun to scratch the surface of this problem. Moreover, the problem appeared sufficiently important in a number of educational contexts to warrant further attention.

Instructional Sensitivity of Tests. The impetus for the work on multilevel approaches to test development and interpretation is the increasing concern about the instructional sensitivity of standardized achievement tests. This concern derives from several aspects of current thinking about such testing. First, there is support for the notion that test performance is high when there is substantial overlap between the content of the test and the content of instruction (e.g., Armbruster et al., 1977; Jenkins & Pany, 1976; Leinhardt & Seewald, 1980; Madaus et al., 1979; Walker & Schaffarzik, 1974). Given this connection, the evidence of wide variation in content coverage in the major standardized achievement tests (Porter et al., 1978) raises the question of whether schools have carefully selected the test which best fits their curriculum (and whether this is even possible in a district with many schools). Second, researchers from diverse viewpoints have argued that while the broad spectrum of standardized achievement tests may be useful indicators for illuminating state and national policies, these tests are insensitive to instructional or program effects (Airasian & Madaus, 1976, 1980; Berliner, 1978; Carver, 1974, 1975; Hanson & Schutz, 1978; Madaus et al., 1979, 1980; Porter et al., 1978).

The weak evidence of schooling and program effects (Averch et al., 1972; Coleman et al., 1966; Stebbins et al., 1977) in the face of strong beliefs that students do learn from given school and program experiences is largely responsible for current challenges to the instructional and program relevance of standardized achievement tests. The challenges from researchers knowledgeable about classroom practices and processes are based on the argument that as long as teachers have the freedom to choose areas of coverage and emphasis, tests cannot be expected to have relevance for all classrooms. Curriculum developers offer similar reasons for suggesting that tests are not appropriate to the content of their curricula. While these arguments have intrinsic merit, they raise as many questions about the appropriateness of instructional coverage decisions by teachers and curriculum developers as they do about the utility of the tests for measuring skills that should be part of the repertoire of the nation's students.

These concerns about the instructional sensitivity and program relevance of norm-referenced achievement tests have caused some educational researchers and practitioners to turn to criterion referenced measurement (e.g., see Berk, 1980; Baker, Linn, & Quellmalz, 1980; Harris, Alkin, & Popham, 1974; Popham, 1978). When looking at a single program with common goals, objectives, and curriculum coverage, criterion-referenced tests can provide a better measure of the quality of instruction when targeted to the specific goals and objectives of the program. However, once a study shifts from a single uniform program to examine multiple groups (e.g., classroom or school) that may share a common general goal but approach it differently (e.g., different specific instructional objectives, different sequencing, or different relative emphasis across objectives), trouble arises in trying to develop criterion-referenced tests, both specific to the program of each group (classroom or school) and yet general enough for comparisons across groups. One alternative is to build criterion-referenced measures that contain all the objectives of all the programs. But this strategy can rapidly become unwieldy because the differences between programs generate too much material to test. Furthermore, when some programs cover more objectives than another, they are still at an advantage because there are fewer novel topics covered on the exam.

Given the problems with using criterion-referenced tests to measure differences between groups which differ in instructional objectives and/or approaches, it is not surprising that norm-referenced tests continue to be used for cross-program (school or classroom) comparisons, especially when they are judged to adequately cover (at least at some level of generality) the common part of the curriculum. The challenge is to insure that whatever measures are used to judge impact are sufficiently sensitive to differences in programs and instructional groups. Since standardized tests are at present the primary evidence for such judgments, the extent to which they perform their desired function warrants attention.

Measuring Programs As Well As Students. There is a perhaps too subtle shift in emphasis implicit in our concerns about the instructional and program relevance of measures of student performance. The rationale for the current investigation might instead be viewed as part of a shift in the conception of the purpose for standardized achievement testing in education. A traditional conception would clearly emphasize obtaining a description (measure) of what students know and how their knowledge compares with that of a relevant group (classmates, same school, same grade level, publishers' norms, etc.). The same rationale holds whether one is talking about norm-referenced or criterion-referenced measurements though with the latter, both the degree of specificity of the pertinent body of knowledge and the nature of the comparison (to a given level of performance within the domain of knowledge reflected in the test) are changed. Measuring what students know is still the primary concern.

This individualistic conception of achievement measurement served well as long as the measures of performance were intended only to help reach decisions about individuals (e.g., Does the student have the necessary background knowledge for Algebra II? Who should be selected for an academic scholarship? Which students need remedial instruction in reading? Should the student be advanced to the next objective or spend additional time on the ones already studied?). While the level of generality required in dividing performance measures into content domains might vary depending on the specific circumstances (see Baker, 1981), that the decisions are being made about individuals is still the dominant feature of this kind of achievement measurement, not whether the tests are norm or criterion referenced.

At a simpler period in our history when American citizens were less mobile and more homogeneous, school "systems" were smaller, fewer students advanced to each higher level of the educational system, and there was less to be learned and a greater consensus (folklore) on instructional content and method, operating by a strictly individualistic conception of achievement measurement may have been the proper role for testing in schools. However, the growth in the diversity of modern American society, with the accompanying expansion of the educational level of the citzenry, the information and knowledge to be learned, the centralization of schools into larger school systems and the broadening of the array of curriculum and instructional alternatives, raises questions about the adequacy of purely individualistic models of achievement testing for meeting the changing organization, operations and needs of American education.

Under present conditions in education, then, it seems particularly appropriate to delineate an additional conception of the purpose of achievement testing. This conception emphasizes the role of performance on achievement tests as measures of the quality of the student's educational experiences. Under this conception, the focus shifts from obtaining a status assessment of the individual student to an examination of whether students coming from given educational programs have obtained certain levels of knowledge. The focus is no longer strictly on the student; the school system through its choice of programs in which to participate, through the curriculum decisions about what to teach, through the specific instructional activities of individual teachers and through the coordination of these activities among teachers (both at the same and at different grade levels or subject matters) in the same school and district is viewed as having a direct responsibility to accomplish its educational goals for its students and is held accountable by the public for its actions.[1] Decisions about programs (e.g., How does the performance of students in the pull-out program compare to performance in mainstreamed instruction with more educational assistance in the classroom? Is the special tutorial program enhancing student learning?) and instruction (e.g., Are students in school (classroom) A showing sufficient educational progress? Are students in classroom A which uses textbook Q learning the same things (and as well) as students in other classes using textbook W? Does the body of knowledge taught students in grade M in school B prepare them adequately for the instruction planned in grade M+1? Which instructional topics need further study to bring students in class (school) P up to an acceptable performance level?) are emphasized in addition to concerns about individual learners.

This conception of testing as a means to examine the results of edcuational programs is in line with the concerns of researchers and policy-makers interested in measuring program and schooling effects. More importantly, we argue that this view of achievement testing is consonant with current emphasis on linking testing and instruction in schools and on systemic efforts at program and instructional improvement. It is also clear that this conception places greater emphasis on the aggregation of test scores across students within classrooms, schools, programs, districts, etc., in order to provide information in a form that is more directly relevant to program and instructional decision-making than strictly student level data would.

Psychometric Considerations. Given a concern for measuring program and instructional differences as well as individual differences, the complaints about the traditional psychometric basis for standardized test construction are well-taken. While these tests have been used to assess the achievement or ability differences among individuals, as well as ranking the achievement differences among aggregates of individuals (e.g., classes or schools), the psychometric model used in test construction has focused primarily upon the former. Some critics have argued that tests designed to differentiate among individuals maximize the within-school differences relative to the between-school or between-program differences (Airasian & Madaus, 1980; Carver, 1974, 1975; Lewy, 1973; Madaus et al., 1980).

Theoretically, of course, there is no reason to assume that a test designed to measure individual differences cannot also measure school or program differences. However, the bulk of the evidence from school effectiveness studies seems to suggest that either school or program differences do not exist or we are measuring the differences improperly (Madaus et al., 1980).

Multilevel Considerations. The concerns cited above seem to reflect the same units of treatment and analysis issues which underly much of the recent work on analysis of multilevel educational data (Barr & Dreeben, 1977, 1981; Burstein, 1980a, 1980b; Cooley, Bond, and Mao, 1981; Cronbach, 1976; Wittrock & Wiley, 1970). Cronbach (1976) directly addressed the units of analysis implications for test construction and interpretation and a few studies (e.g., Airasian & Madaus, 1976; Lewy, 1973; Madaus, Rakow, Kellaghan, & King, 1980; Rakow, Airasian & Madaus, 1978) have sought to use test data from multiple levels to reflect schooling and program effects. These efforts barely hint at the possibilities, however.

We argue that multilevel examinations of test item data have the potential to lead to better informed test development, analysis, interpretation, and reporting procedures. For example, careful investigations of test item data might enable one to identify effects due to background differences (e.g., prior learning, sex, socioeconomic and demographic differences), instructional coverage and emphasis, and instructional organization (e.g., grouping and pacing effects). If these separate effects can be identified, it would then be possible for school personnel to reconstruct from item data, a variety of composites which are potentially sensitive to the context factors of their choosing. Likewise, test developers could include in their test development activities and procedures which would guard against unknowingly selecting items influenced by "irrelevant" context and situational characteristics (where "irrelevancy" is determined by the purposes for which the test would be used). At the least, developers would be better able to describe the properties of their tests after carrying out a multilevel examination of their properties.

Our activities under the present grant period were directed to identifying analytical methods which can distinguish the effects of various factors that affect between-group (class, school) and within-group test performance. It was expected that such a multilevel examination would facilitate the use of test data in program and instructional decision-making at various levels of the educational system. Hopefully, the analytical strategies are equally applicable to tests developed for either norm-referenced or criterion-referenced usage.

Methods

The actual empirical investigation undertaken focused on two general approaches for measuring between-group (classroom, school, program, etc.) differences in test performance. Both approaches consider the empirical characteristics of between-group performance on test items or subsets of test items.

Investigations at a level below the total test are considered essential to detect differences in the content, sequencing, and quality of instruction.

Since one is seldom interested in the consequences of no math instruction (versus some), but is often interested in the choice between time spent on and methods used in developing, say, computational skills, one is likely to miss relevant differences in the effects of instruction by considering only total test scores.

Desirable vs. Available Study Characteristics. The practical scenario that guided our empirical inquiry was an examination of the data from a standardized testing program conduction within a school district.[2] Ideally at any given grade level, these data would be available at the item level for students within a number of classrooms within the district's schools. Under these circumstances, the student responses to individual test items can be both vertically aggregated (instructional groups within classrooms, classrooms within schools, schools within the district) as well as demographic groups (e.g., males vs. females, monolingual vs. bilingual students, different demographic groups), and horizontally aggregated (across items within a narrow domain, to the level of instructional units, at the typical subtest level on achievement tests, as well as specific combinations of subtests and other classifications of items (e.g, according to process being tested, linguistic features, task structure, etc.)) to obtain the desired specificity of information about program and instructional differences. Thus, an investigator would be able to generate indices of the distribution of test performance for a variety of groupings of students (by class, school, ethnic group, etc.) under alternative rules for content classification.

The empirical work was conducted on data from the Beginning Teacher Evaluation Study (BTES; Fisher, Filby, Marliave, Cahen, Dishaw, Moore, & Berliner, 1978). The primary data set contains test performance of 125 fifth-graders (approximately 6 students from each of 22 classrooms) on the fifteen fraction items from the BTES test battery. The fractions subtest was administered on three occasions -- prior to any significant amount of fractions instruction (Occasion B, December), near the end of the school year (Occasion C, May), and again the following October (occasion D). Fractions was chosen because of its predominance in fifth grade mathematics instruction.

The six students in each classroom selected for intensive study, scored between the 30th and 60th percentile on a beginning-of-the-year prediction battery given to all the students from the 22 classrooms. The limitation on the number of students studied was due to the intensive classroom observations (approximately 25 full days during the year) and teacher record keeping requirements. (Teachers were required to keep daily records of the specific time allocated to different content areas for each student in the intensive study.) The students were chosen from the narrower range to ensure that the study concentrated on the learning experiences of "typical fifth graders". In addition to the test information described above, our investigation also included the BTES measures of Allocated Time in fractions between the B and C test occasions, student Engagement Rates during mathematics instruction, and the proportions of student time during math spent on tasks with which they achieved high success (missed very few problems) and low success (answered very few problems correctly). Additional details about the data set are contained in the longer report in Appendix A.

In practice, the BTES data differed in several respects from the data described under the ideal scenario. Typical classrooms have more students and most likely a broader range of abilities. Moreover, the content investigated is much narrower than would be typically available in a standardized test battery though there were perhaps more items devoted to fractions than one would typically find. Moreover, the full sample was more homogeneous than the fifth-grade population as a whole. It might also be the case that mathematics performance levels of the classrooms was more homogeneous than typical distribution of fifth-grade classrooms.

These departures from the ideal both helped and hurt our empirical efforts. The overall sample size was sufficiently small to allow thorough empirical analysis by both statistical and graphical means at reasonable cost. We were better able to trace particularly interesting results back to their source than one could with larger data sets. On the other hand, the small sample restricted the power of the statistical tests one might perform (we were more interested in the magnitude of particular indices rather than their statistical significance) and caused certain empirical indices to be overly sensitive to the atypical performance of individual students within classrooms.

Similarly, the restriction in test content had mixed consequences. On the one hand, we were gratified to find that potentially important differences in instructional activities could be identified by examining class-level performance on items and relatively homogeneous subsets of items. There would seem to be clear advantages in being able to pinpoint instructional effects at a level of specificity suitable for instructional remediation. On the other hand, a broader array of content was never investigated, there is no way to determine whether the methods used are sensitive to instructional and program differences at a higher level of generality. Research by Madaus, Airasian, and their associates and by Harnisch and Linn (1981) does suggest, however, that the methods studied are applicable to data covering a broader range of content.

We will not comment further on the limitations of our empirical work. Clearly, more empirical efforts are needed to determine just how useful multilevel methods can be in test development and interpretation in local school settings.

Specific Analytical Procedures. As stated earlier, our empirical investigation of between-group program and instructional differences emphasized two distinct approaches. In the first approach, the empirical properties of five indices of item discrimination between groups were investigated. The merits of each index as a criterion for selecting items during test construction were explored. Scales were constructed by choosing items that exceeded a certain level on a specific index of between-group item discrimination. The empirical properties of the constructed scales were then examined and compared with the characteristics of the 15-item fractions total score. The five indices investigated were as follows:

(a) the item intraclass correlation (the proportion of variation in item scores associated with between-class sources of variation);

(b) the combination of item intra-class correlations used in conjunction with between-class item intercorrelations (i.e., the correlations of class mean performance on one item with class mean performance on other items);

(c) the between-class correlation of item performance with total test performance (the group-level analogue of the point-biserial correlation);

(d) a discriminant analysis in which items are used to discriminate among classrooms; and,

(e) the between-group correlation of item performance with a measure of instruction (in this case, time allocated to fractions instruction)..

The criteria used to judge the merits of specific indices included the intraclass correlation of the constructed scale, the magnitude of the effects of instructional variables in regression analyses with student performance on the constructed scale as the dependent variable and between-class and within-class instructional and background measures as explanatory variables, and the overall proportion of "variation explained" (R^2) in student performance. The belief was that specific indices would lead to the construction of scales that retained between-group variation in test performance, increased the relationship of instructional variables to performance and required fewer test items.

The second group of analytical strategies involved adapting procedures previously employed for examining patterns of test item responses of individual students to detect differences between groups (classes in this study) of students. Patterns of correct item responses were investigated through the generation of class-level variants of the Student-Problem Chart developed by Sato (1980). The properties of the mean and standard deviation of Sato's caution index (a measure of the anomalousness of an individual's pattern of correct item response) as a possible statistical measure of differential instructional coverage and emphasis across classrooms were also explored. Finally, the use of the patterns of incorrect item responses as information about between-class instructional differences was examined.

Results

Subsets of Group Sensitive Items. The investigation of the five alternative indices for selecting items for constructing scales more sensitive to group differences pointed to a number of similarities and differences among the indices. First, the indices tended to select slightly different subsets of items. Moreover, the items selected by most indices did not represent any clear content clusters, but rather specific empirical nuances that aligned the analytical foundation for a specific index with the characteristics of student performance. Thus, investigators are likely to need to use several indices to avoid basing item selection on special circumstances existing in a given sample of classrooms and schools.

Second, the scales constructed by all five indices exhibited approximately the same proportion of between-class variation (ranging from .42 to .50) as the total scale (.47). This level of retention of variation was obtained despite one-third (10 item) and two-third (5 item) reductions in test length. Obviously, focussing on indices of between-group discrimination accentuates the between-class differences in item performance that was the basis for their consideration in the first place. Unfortunately, the relationships of the scales to the instructional and background variables fluctuated according to the index used for item selection. As might be expected, the index based on the between-class correlation of the items with instructional variables was most effective in building a scale sensitive to the variable used to select items. Other differences were less predictable. The obvious conclusion from the analysis was that if investigators know the variable according to which they wish to distinguish performance, then selecting items on the basis of their relation to that variable is an effective strategy for empirical item selection.

Finally, the stability of the indices was investigated by comparing scales formed using the data already described with the scales formed from a limited set of pilot data (5 full classes containing approximately 120 students). None of the indices of item discrimination between groups were particularly stable across samples. Different items were selected, the intraclass correlations for the constructed scales changed and the relation of the scale to instructional variables fluctuated. However, the limited number of groups in the pilot study might be at least partially responsible for the observed instability.

Patterns of Item Response. The examination of between-class patterns of correct and incorrect item responses indicated that the patterns of responses were related to group membership. Moreover, since results held up even after controlling for between-class differences on the pretest, the pattern of responses appears to be related to instructional coverage and emphasis.

The patterns of correct item response on the posttest clearly showed a relationship to instructional coverage that were not visible prior to instruction. For example, certain classes with only poor or average performance in the addition of fractions, exhibited high performance on the more difficult "algebraic manipulation" topic. The differences in coverage and emphasis turned out to be most evident at the item level. For example, students in some classrooms managed to learn simple addition and subtraction of fractions with common denominators and virtually nothing else.

The results from the use of the class mean and standard deviation on the caution index as statistical indices to detect unusual instructional patterns were mixed. Classrooms whose unusual instructional coverage and emphasis was evident from the patterns of correct responses tended to have high mean caution indices. Unfortunately, there were several classes in which the anomalous response pattern for a single student (out of 6) also resulted in high mean caution indices. However, since these classrooms also tended to exhibit high variability in the caution index, it was still possible to separate classrooms with distinctive instructional patterns from those with variable student response patterns. The confusion of individual with group anomalousness should be even less likely in regular size classes.

The class-level analysis of patterns of incorrect item responses was particularly informative. There were clear instances where students in the same classroom exhibited a common incorrect problem solving procedure (e.g., adding both numerator and denominator in the addition of fractions). The reasons for this incorrect procedure may be traceable to inadequate instruction or simply lack of instruction when the faulty procedure was present prior to instruction. Overall, there was considerable evidence that error patterns reflect both random and systematic processes and that systematic errors have both individual-specific and group-specific determinants.

Concluding Comments

As with any research, the conclusions of this study are limited by the data employed and further research is needed. Nevertheless, the present investigation does provide support for arguments that tests can be constructed in ways which are more or less sensitive to desired group characteristics (e.g., instructional and program differences) and investigations of group-level patterns in test item responses can provide important information about the group-based differences in instructional experiences.

Having concluded that the multilevel approaches to test development and interpretation are potentially beneficial, we need to comment further on the conditions under which we expect these methods to be maximally useful. In order to achieve maximum benefits from procedures for selecting group-sensitive items, it appears that one needs to know the specific characteristics whose between-group effects one wants to measure. For instance, it is logical to choose items which exhibit high relationships to time allocated to instruction if the intended purpose of the scales constructed from the items is to distinguish the consequences (in future samples) of differences in instructional coverage. This is precisely the basis for the item selection procedures employed in the BTES study and might be used in other instances where the intent is to monitor the effects of such instructional differences. The problem is that in many cases, investigators do not know nor are they able to anticipate the characteristics of groups that are most salient to their purposes. Alternatively, the number of characteristics of interest may be large and their interactions may be complex in natural classroom settings. Under these circumstance, the investigator is forced to explore a number of alternatives in the hope of discerning patterns of group sensitivity that reflect on the questions of interest. This is likely be both a time-consuming and difficult task.

We are less concerned that investigation of group-level patterns in test item performance can go awry. In fact, group-level information appears to be particularly well-suited for the purpose of forming decisions about instruction and program effects. We can envision providing teachers (and groups of teachers) with the patterns of performance for their own class as well as patterns for seemingly similar classrooms. While this class-level information may not be sufficiently diagnostic about an individual student's problems, it can potentially pinpoint for teachers (and groups of teachers) the consequences of their particular decisions about instructional coverage, emphasis, and method. As such, class and school level patterns of test item performance would seem to be a valuable element of information-based program improvement activities in individual classrooms, schools, and school districts.

What remains to be determined about investigations of group-level item response patterns is whether these methods become intractible once the number of groups and number of items becomes large. We also need to know more about which special characteristics of groups (e.g., heterogeneity of ability or differential instructional coverage within classrooms) or items (e.g., the diversity of content, information processing requirements) cause examinations of response patterns to be more or less fruitful. There is also a question of how the amount of information and the method of reporting it affects the usefulness of these procedures for specific audiences (e.g., teachers, principals, administrators, evaluators). While the successful results from examinations of graphical procedures is heartening, there are clearly limits on how far one can go before even the simplest form of data display becomes an unintelligible blur for the practitioner.

Given the above concerns, the next phase in this investigation of multilevel methods for test development and interpretation should be obvious. It is time to investigate the utility of these multilevel methods in actual testing and test reporting procedures in schools and school districts. Studies in such contexts are necessary to identify the boundaries of the practical applications of multilevel perspective toward test usage in local school improvement efforts.

FOOTNOTES

(1) We do not intentionally ignore the role of the home in this conception. However, school systems have the responsibility of communicating their educational goals to parents and providing them a means for participat`ng in the education of their children. Moreover, schools cannot abdicate their responsibilities in the development of a well-educated citzenry simply because of shortcomings in the home.

(2) The scenario need not be restricted to the school district level and below, especially when broader curriculum and program evaluation issues are at stake. However, it seems unlikely that the kinds of program and instructional improvements of interest here can be reasonably accomplished through examination of higher-level data except to the extent that a given district judges its performance by comparison with other districts. The form of signal reflected by district-level data is almost invariably at least a step removed from the level where program and instructional changes can be implemented. It is at the school-building level and below where instructional management occurs. Thus, we have concentrated our efforts on methods for using test information at the level of school and classroom. We return to this issue later on.

References

Airasian, P.M., & Madaus, G.F. A study of the sensitivity of school program effectiveness measures. Report submitted to the Carnegie Corporation, New York. Chestnut Hill, MA: Boston College, School of Education, 1976.

Armbruster, B.B., Steven, R.O., & Rosenshine, B. Analyzing content coverage and emphasis: A study of three curricula and two sets. Technical Report No. 26, Center for the Study of Reading, University of Illinois, Urbana-Champaign, 1977.

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, CA: The Rand Corporation, 1972.

Baker, E., Linn, R.L., & Quellmalz, E. Knowledge synthesis: Criterion-referenced measurement. Center for the Study of Evaluation, University of California, Los Angeles, CA, 1980.

Barr, R., & Dreeben, R. Instruction in classrooms. In L.S. Schulman (Ed.), Review of research in education (Vol. 5). Itasca, IL: Peacock, 1977.

Barr, R., & Dreeben, R. How schools work: A study of reading instruction. Draft Manuscript, 1981.

Berk, R.A. (Ed.) Criterion-referenced measurement: The state of the art. Baltimore, MD: The John Hopkins University Press, 1980.

Berliner, D.C. Studying instruction in the elementary school classroom: Clinical educational psychology and clinical economics. Paper commissioned by the Education, Finance, and Productivity Center, Department of Education, University of Chicago, 1978.

Burstein, L. The roles of levels of analysis in the specification of educational effects. In R. Dreeban & J.A. Thomas (Eds.), The analysis of educational productivity, Vol. I: Issues in mirco-analysis. Cambridge, MA: Ballinger, 1980, 119-190. (a)

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.: AERA, 1980, 158-233. (b)

Carver, R.P. Two dimensions of tests: Psychometric and edumetric. American Psychologist, 1974, 29, 512-518.

Carver, R.P. The Coleman Report: Using inappropriately designed achievement tests. American Educational Research Journal, 1975, 12, 77-86.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, S., Weinfeld, F.D., & York, R.L. Equality of educational opportunity. (2 Vols.). Office of Education and Welfare. Washington, D.C.: U.S. Government Printing Office, 1966.

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: John Hopkins University Press, 1981.

Hanson, R.A., & Schutz, R.E. A new understanding of schooling effects derived from programmatic research and development. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, April 1978.

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), 133-146.

Harris, C.W., Alkin, M.C., & Popham, W.J. (Eds.) Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles, CA: Center for the Study of Evaluation, 1974.

Jenkins, J.R., & Pany, D. Curriculum biases in reading achievement tests. Technical Report No. 16, Center for the Study of Reading, University of Illinois, Urbana-Champaign, 1976.

Leinhardt, G., & Seewald, A. Overlap: What's tested, what's taught? Journal of Educational Measurement, 1981, 18, 85-96.

Lewy, A. Discrimination among individuals vs. discrimination among groups. Journal of Educational Measurement, 1973, 10, 19-24.

Madaus, G.F., & Airasian, P.W. The measurement of school outcomes in studies of differential school and program effectiveness. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA, 1980.

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

Madaus, G.F., Kellaghan, T., Rakow, E.A., & King, D.J. The sensitivity of measures of school effectiveness. Harvard Educational Review, 1979, 49, 207-230.

Popham, W.J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1978.

Porter, A.C., Schmidt, W.H., Floden, R.E., & Freeman, D.J. Practical significance in program evaluation. American Educational Research Journal, 1978, 15, 529-539.

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

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

Stebbins, L.B., St. Pierre, R.G., Proper, E.C., Andersen, R.B., & Cerva, T.R. Education as experimentation: A planned variation model. Cambridge, MA: Abt Associates, 1977.

Walker, D.F., & Schaffarzik, J. Comparing curricula. Review of Educational Research, 1974, 44, 83-111.

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

APPENDIX A

DRAFT

Deliverable November 1981

Empirical Studies of Multilevel Approaches to Test Development and Interpretation: Measuring Between-Group Differences in Instruction

M. David Miller

Grant Number

NIE-G-80-0112

CENTER FOR THE STUDY OF EVALUATION
GRADUATE SCHOOL OF EDUCATION
University of California, Los Angeles

TABLE OF CONTENTS

Abstract

Chapter Page

1 INTRODUCTION 1

Test Construction 6

Analysis and Interpretation of Test Item Data 8

Overview of Later Chapters 11

2 LITERATURE REVIEW 13

School Effects Concerns 13

Test Construction 20

Analysis and Interpretation of Patterns of Item Response 28

3 METHODS 40

Analytic Strategies 40

Data Base 47

Data Analysis 56

4 SUBSETS OF GROUP SENSITIVE ITEMS 66

Descriptive Statistics 67

Phase III-B 67

Phase III-A Continuation 82

Analyzing Subsets of Items Sensitive to Group Differences 89

Summary 112

Table of Contents (Continued)

CHAPTER — PAGE

5 ANALYSIS OF PATTERNS OF ITEM RESPONSE 117

Analysis of Patterns of Correct Item Response 117

Patterns of Error Response 135

Summary 149

6 SUMMARY AND CONCLUSIONS 151

Overview of Chapters 151

Summary of Results 153

Suggestions for Further Research 158

Concluding Remarks 161

Footnotes 164

Appendices

A - Fraction Items - Phase III-B 165

B - Fraction Items - Phase III-A Continuation 169

C - Zero Order Correlations of Constructed Scales 173

D - Caution Index and Student Response Patterns 178

References 191

ABSTRACT

The largely negative results of the studies of the effects of schooling suggest that the relationship of school or program inputs to student achievement is negligible or nonexistent after controlling for home background and entering ability. Because of a belief that schooling does affect student achievement, researchers have questioned the empirical and measurement techniques used to evaluate the effects of schooling on student achievement. One area of concern is with the instructional sensitivity and program relevance of standardized norm-referenced achievement tests used in these studies.

Since education occurs in a multilevel system (e.g., students within classes, classes within schools, etc.), one possible shortcomming of the major standardized norm-referenced achievement tests is their failure to take into account the nesting of units in the educational system during test construction analysis, and interpretation. Consequently, this study uses item data at the class level to construct tests potentially more sensitive to instructional differences and to examine patterns of item response which may help describe instructional differences.

Data from the fractions test from the Beginning Teacher Evaluation Study, was used in an empirical examination of test construction and the analysis of patterns of item response at the class level. In the construction phase of this study, it was found that selecting items from an index of discrimination between groups did lead to

scales more sensitive to instructional differences. Moreover, specific indices of item discrimination between groups led to scales with different properties which might be useful in different contexts.

From the analysis of patterns of item response, it was found that not only did patterns of correct and incorrect item response vary as a function of class membership, but that the patterns of response reflect substantively meaningful differences in instruction. The patterns of correct item response clearly showed a relationship to instructional coverage and emphasis that were not visible prior to instruction. The patterns of incorrect item response were also found to depend on class membership.

Overall, the results suggest that important group differences, including instructional and program differences, can be gleaned from a group-level analysis of item data. Furthermore, the total test score will often mask these important differences.

CHAPTER 1

INTRODUCTION

For much of the past 15 years, studies of the effects of schooling and large-scale evaluations of educational interventions have indicated that the relationship of school or program inputs (e.g., per pupil expenditures, number of aides, or differential allocation of time in a given subject area) to pupil outcomes is negligible or nonexistent after controlling for home background and entering ability (Averch, Carroll, Donaldson, Kiesling, & Pincus, 1972; Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, & York, 1966; Comber & Keeves, 1973; Jencks, Smith, Acland, Bane, Cohen, Gintis, Heyns, & Michelson, 1972; Purves, 1973; Stebbins, St. Pierre, Proper, Andersen, & Cuerva, 1977; Thorndike, 1973). However, a prevailing belief that schooling does affect student achievement and that program and instructional differences are reflected in student achievement has caused many researchers to question the traditional empirical and measurement procedures used to arrive at these conclusions.

The traditional model used in program evaluation and school effects studies is to treat student achievement (usually measured by norm-referenced achievement tests) as a function of home background or socioeconomic status, and school or program inputs (e.g., Coleman et al., 1966; Comber & Keeves, 1973; Purves, 1973; Thorndike, 1973). While it would be preferable to use a pretest-posttest design (Campbell & Stanley, 1963) with the schooling variables equated with the treatment and socioeconomic background entered as an additional covariate, the cross-sectional nature of school effects data often precludes the use of a pretest.

The model used to evaluate the effects of school or program variables on student achievement has been criticized for a number of different shortcomings. Analytical concerns typically focused on problems in the model specification (Bowles & Levin, 1968; Hanushek & Kain, 1972), and in the treatment of multicollinearity between background and schooling effects (Carver, 1975; Mayeske, Okada, Cohen, Beaton, & Wisler, 1973; Mayeske, Wisler, Beaton, Weinfeld, Cohen, Okada, Proshek, & Tabler, 1972). In addition, there has been a lot of criticism of the adequacy of the measurement of both school and program components (House, Glass, McLean, & Walker, 1978; Veldman & Brophy, 1974), and outcomes (Airasian & Madaus, 1976; Haney, 1977; House et al., 1978). Concerns about the adequacy of the measurement of outcomes are of primary interest to this inquiry.

The outcome measures typically used in large-scale program evaluations and school effects studies are standardized norm-referenced achievement tests. These tests, such as the Metropolitan Achievement Test or the California Test of Basic Skills, usually consist of two sets of subtests -- verbal and quantitative. The verbal subtests cover areas of language skills such as spelling, word knowledge, language (grammar), and reading (answering questions about a paragraph), and in the later grades, verbal tests also cover substantive areas such as history and science. On the other hand, quantitative subtests might cover either different areas of mathematics skills (e.g., addition, subtraction, decimals), or the subtests might reflect a more general division into mathematics computation, mathematics concepts (e.g., inequality or time measurement), and solving word problems.

Because norm-referenced achievement tests are typically used to evaluate the effects of schooling, it is not surprising that much of the criticism of the school effects literature has centered around the instructional sensitivity and program relevance of tests (Airasian & Madaus, 1976; Carver, 1974; Porter, Schmidt, Floden, & Freeman, 1978). This concern derives from several aspects of current thinking. First, there is support for the notion that test performance is high when there is substantial overlap between the content of the test and the content of instruction (e.g., Armbruster, Steven, & Rosenshine, 1977; Jenkins & Pany, 1976; Leinhardt & Seewald, 1980; Madaus, Kellaghan, Rakow, & King, 1979; Walker & Schaffarzik, 1974). Given this connection, the evidence of wide variation in content coverage in the major standardized tests (Porter et al., 1978) raises the question of whether schools have carefully selected the test which best fits their curriculum and whether this is even possible in a district with many schools. Second, researchers from diverse viewpoints have argued that while the broad spectrum of standardized achievement tests may be useful indicators for illuminating state and national policies, these tests are insensitive to instructional and program effects (Airasian & Madaus, 1976; Berliner, 1978; Carver, 1974, 1975; Hanson & Schutz, 1978; Madaus & Airasian, 1980; Madaus, Airasian, & Kellaghan, 1980; Madaus et al., 1979; Porter et al., 1978).

These concerns about the instructional sensitivity and program relevance of norm-referenced achievement tests have caused some educational researchers and practitioners to turn to criterion-referenced measurement (for review, see Berk, 1980; Harris, Alkin, & Popham, 1974; Popham, 1978). When looking at a single program with common goals, objectives, and curriculum coverage, criterion-referenced tests can provide a better measure of the quality of instruction when targeted to the specific goals and objectives of the program. However, once a study shifts fro: a single uniform program to examine multiple groups (e.g., classroom or school) that may share a common general goal but approach it differently (e.g., different specific instructional objectives, different sequencing, or different relative emphasis across objectives), trouble arises in trying to develop criterion-referenced tests, both specific to the program of each group (classroom or school) and yet general enough for comparisons across groups. One alternative is to build criterion-referenced measures that contain all the objectives of all the programs. But this strategy can rapidly become unwieldy because the differences between programs generate too much material to test. Furthermore, when some programs cover more objectives than another, they are still at an advantage because there are fewer novel topics covered on the exam. Given the problems in developing and using criterion-referenced tests to measure differences between groups which differ in instructional objectives and/or approaches, it is not surprising that norm-referenced tests continue to be used for cross-program (school or classroom) comparisons, especially when they are judged to adequately cover, at least at some level of generality, the common part of the curriculum.

Since norm-referenced tests will likely continue to be a standard part of school testing programs, it is important to develop methods for improving their program and instructional relevance. In this study, the instructional sensitivity of norm-referenced achievement tests will be examined from an empirical perspective. First, the empirical properties of test items will be used to build instruments more sensitive to school effects. In addition, empirical methods of examining patterns of item responses to gain information about classroom, school, or program differences will be investigated. We will elaborate on our methods for investigation in later chapters. At this point, however, we provide a brief overview of the two main lines of our inquiry into the program relevance and instructional sensitivity of achievement tests.

Test Construction

Depending on the intent of a study, educational investigations focus on educational performance at various levels (individual, classroom, school, etc.) of the educational system. That is, one might be interested in measuring differences among individuals as well as among groups of individuals, such as classrooms, schools, and programs. Because of their pervasive use in educational investigations, standardized norm-referenced achievement tests are often used to assess outcomes at different levels of aggregation. However, the psychometric model used in the construction of such tests is concerned exclusively with one level -- the individual. Yet, focusing on the individual has been shown to yield a different set of items than constructing tests with a focus on the group (Lewy, 1973).

Because instructional (e.g., instructional organization, style, and emphasis) and other program characteristics (e.g., aides, and money) are often administered, and consequently can be measured, only at a level other than the individual (e.g., classroom, or school), it may prove useful to attempt to construct tests that are more sensitive to differences between groups rather than differences between individuals. This might be operationalized by examining the empirical characteristics of item data at the group level. That is, indices of item discrimination between groups might be used as a criteria for item selection during the test construction phase of a study. In this study, a number of indices of item discrimination between groups will be used to construct scales and the empirical properties of the constructed scales will be examined.

Analysis and Interpretation of Test Item Data

In school effects studies and large-scale evaluations, the dependent measure typically consists of a composite of many different skills (e.g.; verbal ability, instead of spelling, grammar, etc.) or multiple composites of skills (e.g., spelling, grammar, etc.). That is, some global measures of achievement such as nonverbal skills, or subsets of skills such as multiplication and division are typically used as outcomes. However, even at the subtest level (e.g., multiplication), differences that exist between classes may be masked by the use of a composite score. As Airasian and Madaus (1976) point out, items may differentiate between groups in different directions, so that when summed, the composite fails to find group differences. For example, given two groups of equal size, if everyone in one group answered one item correctly and a second item incorrectly, while the reverse was true for the other group, then the two items would discriminate perfectly between the two groups individually, but the sum of the two items would show no differences between the two groups.

This masking of achievement differences in tests and subtests is especially important in considering the instructional sensitivity of an achievement measure. Instruction varies across classrooms with respect to topics that are not always defined at the test or subtest level. Instruction can vary in the amount of time devoted to a given topic, the approach taken to a given topic, the quality of time devoted to a given topic, and the sequence in which topics are covered. Analyses of test item data need to be sensitive to these differences. Since achievement differences might be manifested in the order, thoroughness, and quality of coverage of specific topics, performance on subtopics of an achievement test should vary not only on the basis of individual abilities, but also as a result of instructional differences. Different instructional settings might lead to different difficulty hierarchies for items.

In order to detect these differences, test data need to be examined at the item or the subtopic level. Consequently, this study will examine patterns of item response to try to gain information about instructional differences that might not be gleaned from the total test or subtest score. To facilitate the examination of patterns of item response for group differences, certain strategies for analyzing patterns of response at the individual and group level will be employed.

In addition to examining correct item response patterns and their relationship to instructional process,the possibility that an incorrect response might be related to instruction will be examined. The last decade has seen a growth in the literature which suggests that errors on achievement tests are often not random, but are the result of an incorrect algorithm (or procedure). A student's incorrect response may be a function of a number of different variables. For example, the student may have carried forward an incorrect algorithm from some lower skill (e.g., an incorrect algorithm for addition repeating itself in multiplication), or an incorrect algorithm may have developed because the student was absent on the day that the skill was being taught. Alternatively, the incorrect response pattern may be related to the instruction received. That is, an incorrect procedure shared by the majority of the students in a classroom is probably influenced by the instruction in the given content area. The teacher may be unknowingly or inadvertently teaching an incorrect procedure, or something in the instruction may be encouraging the transfer of an incorrect algorithm from another area. Regardless of the reason for any recurring incorrect response common to members of the same group, useful information can be gained about what errors the students are making and what adjustments to instruction need to be made to correct the problem.

Overview of Later Chapters

Earlier literature on school effects and concerns about the instructional sensitivity and program relevance of standardized norm-referenced achievement tests are reviewed in Chapter 2. Also the pertinence of a multilevel perspective to this investigation [illegible] discussed. Next earlier attempts to construct more sensitive indicators of group differences are reported. Finally, work on the analysis of differences in patterns of correct and incorrect responses at the individual level which might be applicable at the group level are reviewed.

In Chapter 3, the empirical techniques which might be useful in measuring between-group differences in test item performance are outlined. Then the data base which will be used to apply the empirical techniques is described. The chapter concludes with an outline of the procedures to be used in applying the empirical methods to the data base.

In Chapters 4 and 5, the data analyses outlined in Chapter 3 are reported. Chapter 4 begins by describing the empirical characteristics of the data base. Next Chapter 4 will examine the empirical properties of the items which will be used to form more sensitive indicators of group differances, as well as the scales which are formed. In Chapter 5, the use of group-level patterns of both correct and incorrect responses to gain information about classroom differences is explored.

Finally, Chapter 6 will attempt to summarize the results of Chapters 4 and 5. In addition, future implications of this study and topics warranting further attention will be discussed.

CHAPTER 2

LITERATURE REVIEW

Two lines of inquiry might be followed to improve the instructional sensitivity and program relevance of norm-referenced achievement tests. One approach is to reconsider the empirical properties used in constructing such tests (e.g., Airasian & Madaus, 1976; Lewy, 1973). A second approach is to use patterns of item response after testing to gain information about differences between groups. This chapter will begin with a brief review of the background work on how the question of measuring school effects became an issue and why the notions from the literature on the analysis of hierarchical or multilevel data seem pertinent. Second, past efforts at building more sensitive indicators of group differences through the use of indices of item discrimination between groups will be discussed. Finally, procedures for the analyses of patterns of item response which might be useful at the group level will be described.

School Effects Concerns

The largely negative results from the studies of the effects of schooling (e.g., Averch et al., 1972; Coleman et al., 1966; Comber & Keeves, 1973; Jencks et al., 1972;

Purves, 1973; Stebbins et al., 1977; Thorndike, 1973) have been disputed for various reasons. Economists have argued that the analytical methods used were inappropriate for determining the effects of school resources on student achievement (Bowles & Levin, 1968; Hanushek, 1970; Hanushek & Kain, 1972; Levin, 1970). Also, based on their reanalysis of the Coleman Report, Mayeske and his associates (Mayeske, et al., 1972, 1973) argued that the effects due to schools were much larger, when the variance shared by home background and school inputs was not attributed to home background. Other sources of concerns included the adequacy of the measurement of educational process (Veldman & Brophy, 1974) and achievement (Airasian & Madaus, 1976).

While researchers have recognized the inherent problems in using questionnaire data to measure schooling (Dyer, 1969; Jencks, 1972; Veldman & Brophy, 1974), the majority of the measurement criticism has centered around the instructional sensitivity and program relevance of standardized norm-referenced achievement tests which typically serve as measures of the outcomes of schooling (Airasian & Madaus, 1976; Berliner, 1978; Carver, 1974; Hanson & Schutz, 1978; Leinhardt & Seewald, 1981; Madaus, et al., 1979, 1980; Porter, et al., 1978). This concern has stemmed in part from the evidence that there is wide variation in the content covered in the major standardized achievement tests (Porter et al., 1978), and that achievement is higher when the overlap of test content and instructional content is greater (Armbruster, et al., 1977; Jenkins & Pany, 1976; Leinhardt & Seewald, 1981; Madaus et al., 1979; Walker & Schaffarzik, 1974).

Given that groups (e.g., schools, classes) differ in objectives and approaches to those objectives, it seems clear that different instruments would overlap to different degrees with each group's coverage of the material tested. This phenomenon has raised the question of whether groups with diverse objectives should be tested with different instruments that are sensitive to their respective objectives or whether a common set of instruments should be applied to all groups (Ellett, Haun, Pool, & Smock, 1979; Madaus et al., 1980; Rivlin & Timpane, 1975; Wargo & Green, 1978; Weikart & Banet, 1975). Since one purpose of the school effects research is to compare performance across groups (i.e., classrooms, schools, school districts, etc.), it is assumed that some measure of the common objectives of the groups is desirable. In addition, some measures specific to the individual groups might be used in conjunction with the common measures. In that way, groups could be compared on common objectives as well as group specific objectives

(Rivlin & Timpane, 1975). However, even on a common measure, there will be group differences. Consequently, analytical concerns arise about how to best describe the differences between groups.

Multilevel Analysis

Once one accepts the notion that the measurement of instructional and program effects rather than strictly individual differences is of interest, one is confronted with the fact that education takes place in a multilevel system. That is, students are nested within classes; classes are nested within schools; and so on. Furthermore, most basic instructional and program variation occurs at some level higher than the individual. Burstein (1980a), in reviewing three prior studies (Baker, 1976; Murnane, 1975; Wiley & Bock, 1967), concluded that significant effects occured at both the school and class level. In addition, the effects at the two levels were related to subject area. "Mathematics instruction exhibits stronger effects than reading ... with most of the impact associated with classroom-to-classroom differences. This latter finding is as expected. While home influences have a stronger impact on reading, mathematics is a subject matter largely learned in school" (Burstein, 1980a, pp. 142-143).

Substantial variation in instructional and program effects occurs at some level higher than the individual regardless of how the within-class instruction is actually structured. As long as different teachers, or schools, use different means of structuring (e.g., whole group, small group, or individual instruction) or cover different topics or even cover the same topics with different degrees of emphasis and quality, then class or school differences provide potentially detectable variation in instructional treatment that should be manifested in test performance (Barr & Dreeben, 1977; Brown & Saks, 1980; Wiley, 1970).

Given that instructional and program effects occur at a level other than the individual, problems in data analysis can occur. The wide body of literature on multilevel issues in large-scale evaluations and school effects studies (e.g., Burstein, 1980b; Cooley, Bond, & Mao, 1981; Cronbach, 1976; Roberts & Burstein, 1980; Wiley, 1970) has identified a number of concerns which warrant further research. First, analyses can be conducted at various levels -- both within level and between level. Second, the level of analysis does matter because analyses at different levels yield different results. Moreover, the different results may be a direct result of different substantive phenomena (Burstein, 1978, 1980b; Burstein, Fischer, & Miller, 1980; Cronbach, 1976; Cronbach & Webb, 1975). Thus, analyses should be conducted at the level or levels that best fit the substantive model (Burstein, 1980b). This issue of analyzing data at the level specified by the substantive model has been recognized in large-scale evaluations, such as Follow Through (Haney, 1974, 1977, 1980) and the National Day Care Study (Singer & Goodrich, 1979).

While there has been a rapid rise in the concern for multilevel issues in the analysis of large-scale evaluations and school effects studies, most researchers have neglected the impact of multilevel issues on the handling of norm-referenced achievement test data, both in their construction and in the analysis of test item data. A notable exception is Cronbach's monograph (1976, pp. 9.19-9.10) on multilevel issues where he briefly discusses the possible utility of multilevel item analysis and test construction:

> Once the question of units is raised, all empirical test construction and item-analysis procedures need to be reconsidered. Is it better to retain items that correlate across classes? Or items that correlate within classes? A correlation based on deviation scores within classes indicates whether students who comprehend one point better than most

> students also comprehended the second point better than most -- instruction being held constant. A correlation between classes indicates whether a class that learned one thing learned another, but this depends first and foremost on what teachers assigned and emphasized. It is the items teachers give different weight to that have the greatest variance across classes. This (differential emphasis) leads us to regard the between-group and within-group correlations of items as conveying different information, and makes the overall correlation for classes pooled an uninterpretable blend.

Cronbach's comments are revealing because the issue of multilevel analysis is seen in the context of test item data. The analysis of item data at different levels is considered to yield different substantive information. Analysis of item data at the class level is considered to render information about differential coverage and emphasis. Thus, when interested in measuring instructional differences, it may be better to analyze item data at the group level, both in test construction and item analysis.

Test Construction

Historically, empirical analyses of test item data from standardized tests in the test construction phase have been conducted on individual-level data. Typical of most writings on test construction, Henrysson (1971) emphasizes the discrimination between students without reference to the possibility of discrimination at some higher level (e.g., class, school, or program). "One of the main purposes of the (item) tryout is to ascertain to what extent each item discriminates between good and poor <u>students</u> as defined by a criterion. In selecting the criterion to be used, one wishes to find a good measure of the ability or skill the test is designed to assess. Ideally, the criterion should be independent of the item being evaluated.... However, most often the total score on the test itself is used as the criterion" (Henrysson, 1971, p. 135, emphasis added).

Most measurement texts and selected readings never consider the issue of discrimination at any level except the individual (e.g., Anastasi, 1976; Cronbach, 1970; Mehrens & Ebel, 1967; Thorndike, 1971). Even when the primary purpose of a study is to measure differences between groups, tests are still constructed using individual-level data (e.g., Filby & Dishaw, 1975, 1976).

Some authors have argued that tests designed to differentiate between individuals can maximize the within-school differences relative to the between-school differences (Carver, 1974; Lewy, 1973), when the opposite effect may be desired.

Theoretically, of course, there is no reason to assume that developing a test at the individual level will maximize either the between-group differences or the within-group differences. In fact, the correlation of two variables at the individual level is a weighted combination of their between-group correlation and their pooled within-group correlation (see, e.g., Alker, 1969; Hannan, 1971; Knapp, 1977; Robinson, 1950):

$$\rho_{XY} = \eta_X \eta_Y \rho_{\bar{X}\bar{Y}} + \sqrt{1 - \eta_X^2}\,\sqrt{1 - \eta_Y^2}\,\rho_{(X-\bar{Y})(Y-\bar{Y})}.$$

where ρ_{XY}, $\rho_{\bar{X}\bar{Y}}$, and $\rho_{(X-\bar{X})(Y-\bar{Y})}$ are the individual-level, weighted group-level (weighted by group size), and the pooled within-group correlations of X and Y, respectively; and η_X^2 and η_Y^2 are the proportions of variation in X and Y, respectively, that are attributable to group differences (i.e., the correlation ratio or the between-group sums of squares divided by the within-group sums of squares plus the between-group sum of squares). Also, the simple individual-level regression coefficient can be similarly decomposed into the weighted group-level regression coefficient (weighted by group size) and the pooled within-group regression coefficient (Duncan, Cuzzort, & Duncan, 1961):

$$\beta_T = \beta_B \eta_X^2 + \beta_W(1 - \eta_X^2)$$

where β_T is the regression coefficient for regressing the individual-level dependent variable (Y) on the individual-level independent variable (X); β_B is the weighted between-group regression coefficient; β_W is the regression coefficient for regressing the deviations from the group means on Y on the deviations from the group means on X; and η_X^2 is as was defined above.

Although the individual-level relationship could be reflecting differences between groups or differences within groups, the bulk of the school effectiveness literature suggests that school or program differences are small or do not exist after controlling for home background and entering ability (Madaus et al., 1980). However it may simply be the case that between-group differences are not being measured properly.

Some investigators (e.g., Airasian & Madaus, 1976; Lewy, 1973) have argued that the group structure of the data should be taken into account at the test construction phase if test performance is to reflect between-group differences. These investigators have proposed that indices of how well an item discriminates between groups be used in lieu of traditional individual-level indices of item discrimination. In other words, an index of how well item performance varies across groups has been used for including or excluding items from a test instead of the traditional indices of discrimination applied to individual-level data.

The usual approach to standardized test construction (e.g., Anastasi, 1976; Cronbach, 1970; Henrysson, 1971) is to administer a large number of items which are homogeneous in content to a tryout sample. Then a smaller subset of items is chosen on the basis of individual-level indices of item discrimination. This is accomplished by selecting items with a high or moderate discrimination index, usually a point biserial correlation of the total test with the binary scored item (Henrysson, 1971). In addition, item difficulty is used in test construction. Selecting items with a high or low difficulty will yield a test with a smaller variance than a test selecting items with mid-range difficulty. Consequently, a test made of items with a low or high difficulty will narrow the test range and make discrimination more difficult. In addition, the test will require more items for a high reliability when items have a low or high difficulty. Finally, Cronbach and Warrington (1952) found that decreasing the variance of item difficulties leads to a higher test validity.

Items selected for their individual-level discrimination and difficulty may discriminate between groups, but the opposite may result also. Two items that seem the same at the individual level may behave differently at the group level. For example, one item, with a high individual-level discrimination index and a difficulty of .5, may be answered correctly by everyone in half the groups and incorrectly by everyone in the other groups. In contrast, a second item, also with a high individual-level discrimination index and a difficulty of .5, may be answered correctly by half of the individuals in every group. So two items which appear the same at the individual level behave quite differently at the group level. The first item discriminates well between groups, while the second item does not show any group differences. Thus, in order to build a test which discriminates between groups as well as having good individual-level properties, an index of how well an item discriminates between groups would seem to be needed.

Intraclass Correlation

In one of the earliest investigations of instructional sensitivity, Lewy (1973) suggested that the intraclass correlation be used as an index of group discrimination to select items for a test. The intraclass correlation is the proportion of variation in a variable that is attributable to group differences.

The intraclass correlation is the ratio of the variation (or sums of squares) between groups (SSB) to the between-group variation plus the within-group variation (SSW), or the individual-level variation (SST). Thus, the intraclass correlation coefficient equals one when all scores within each group are identical and the only variation is due to differences between groups (i.e., SSB=SST and SSW=0). Conversely, the intraclass correlation coefficient equals zero when all the group means are equal and the only variation is due to differences within groups (i.e, SSW=SST and SSB=0). Lewy proposed that the intraclass coefficient be used to identify subsets of items that maximize the variation between groups on the subscale relative to the individual-level variation on the subscale. When analyzing a fourth grade arithmetic test administered to 3,042 students in 107 classes (Lewy & Chen, 1971), Lewy (1973) found that (1) different items are selected using the intraclass correlation than using traditional measures of discrimination, and (2) two tests with similar item difficulties and item-total correlations, but different intraclass correlations led to different shaped distributions of the class means on the total test (bimodal for high intraclass correlations, unimodal for low intraclass correlations). Lewy did not examine any of the empirical properties of the formed scale, either within the sample or by cross-validation.

Intraclass Correlation - Between-Group Correlations

While the intraclass correlation coefficient may be a useful index of how well an item discriminates between groups, using the index as the sole criterion for item selection may be overly simplistic. As Airasian and Madaus (1976) point out, items that differentiate between groups may do so in different directions. Two items with an intraclass correlation equal to one and a tetrachoric correlation (Divgi, 1979; Lord & Novick, 1968), or the correlation of one item with the other, equal to negative one might have an intraclass correlation equal to zero when summed. Thus, the correlation of the group means on the items might be used in conjunction with the intraclass correlation to select items which reflect between-group variation in the same direction. Airasian and Madaus called this method the intraclass correlation - between-group correlation technique. Items were first selected on the basis of some cutoff on the intraclass correlation. Then, one or more groups of items were formed on the basis of the between-group correlations.

Airasian and Madaus (1976) examined four criteria for item selection. The first two were discussed above -- the intraclass correlation and the intraclass correlation - between-group correlation. In addition, they used a principle components analysis, based on the school-level item difficulties, and a discriminant analysis, where the student scores were used to maximally discriminate between groups defined by school membership. Finally, a discriminant analysis was conducted on those items that also met some cutoff on their intraclass correlation.

Each of the analytic techniques was applied to five different data sets, including both norm-referenced and criterion-referenced tests in reading and mathematics for grades four, six, and seven. Subtests were identified which distinguished between schools within the same sample. The intraclass correlation - between-group correlation was judged to be the best technique for two reasons. First, this technique produced longer subtests, containing more items from the total test, resulting in higher reliability (Kuder & Richardson, 1937; Stanley, 1971). Second, this method led to subtests with a higher percentage of between-school variation than the other techniques.

Because of the success of the intraclass correlation - between-group correlation technique on the sample where the items were selected, a cross-validation was done on four of the data sets. Results indicated that similar scales were obtained when subtests defined on one randomly partitioned group of schools were used to analyze the remaining partition of schools. The same items differentiated performance between groups in two different random partitions. When schools were matched (same teachers for two class sessions), the results were not as clear. The cross-validation showed similar results in two of the four studies for matched schools.

Analysis and Interpretation of Patterns of Item Response

While Airasian and Madaus (1976) did cross-validate one of the techniques used to build subtests, the primary focus of the study was to select groups of items that differentiated between groups better than the total test post hoc. That is, the question they wanted to answer was whether "the use of the total test score in analysis masks significant differences between schools or programs which appeared on subsets of items from within the test" (p.253). Airasian and Madaus found that subtests could be identified that exhibited a higher proportion of between school variation than the total test score. Consequently, they concluded that the "use of the total test score index in school comparisons hides unique and statistically significant school achievement differences at the item or objective level" (p. 259).

Even at the subtest level, however, information is being hidden that might be useful in measuring differences between groups, as well as how an individual is doing. That is, there might still be value to studying patterns of test response below the subtest level, perhaps at the level of individual test items. Moreover, one might detect instructional differences by analyzing the errors made within different groups, as well as the correct responses made. The use of both the pattern of correct responses and the pattern of errors, to provide information about individual differences in achievement, have been examined. Each will be discussed below, including a discussion of how these techniques might give information about groups, as well as individuals.

Patterns of Correct Item Response

While subtests provide information that is not available from the total test, differences exist within a subtest that are being masked by the use of subtests. Subtests in most major standardized tests cover several objectives, domains, and instructional topics. That is, within a subtest, a further subdivision of content structure can often be identified. For example, a subtraction subtest might include four different types of items -- single column subtraction or multiple column subtraction, both with and without borrowing. Given a within-subtest content structure and that between-group differences exist in pacing, sequencing, emphasis, quality of instruction, and so on, patterns of performance unique to different instructional groups or programs can potentially emerge which may not be detectable at the subtest level.

If the instructional process (curriculum content, sequencing, emphasis, and so on) in each instruction group can be described, it should be possible to partition classes a priori into instructional groups with similar item response patterns. However, in the absence of information about instructional process differences, the question still remains whether differences in instructional experiences can be detected by analyzing student item response data. In other words, can differences in instructional process be identified from an examination of post instruction test performance? Assuming this can be done, the question remains of what empirical techniques can be used to accomplish this task. One possible solution may be to apply techniques used in investigating individual-level performance to group-level data (see below).

One potentially useful body of work for examining item response data can be found in the psychometric research in Japan being done by a group of engineers affiliated with the Institute of Electronics and Communication Engineers of Japan (see Tatsuoka, 1979 for a review). Because of the engineering influence, educational research in Japan is often not found in the mainstream of psychometrics. One approach, applied widely in Japan, is the use of item and student response patterns to analyze items, tests, and the students (Sato, 1980; Sato & Kurata, 1977).

Sato's approach to the analysis of item data and student response patterns requires that the data first be arranged into a Student-Problem Chart (S-P). The S-P chart is a matrix of student item responses (1 if the item is answered correctly, 0 if incorrect). The matrix is permuted so that the students (rows) are arranged from top to bottom in the descending order of their total test scores, and the items (columns) are arranged from left to right in the descending order of their difficulties. A hypothetical S-P Chart taken from Harnisch and Linn (1981) is presented in Table 1.

Once the S-P Chart is formed, there are a number of

Table 1

S-P Table for 18 Examinees and 5 Items
(Hypothetical Example)

								Sato's
				Item			Examinee	Caution
Examinee				j			Total	Index
i		1	2	3	4	5	$n_{i.}$	c_i
1		1	1	1	1	0	4	.00
2		1	1	1	0	1	4	.65
3		1	1	1	0	0	3	.00
4		1	1	0	1	0	3	.16
5		1	1	0	0	1	3	.65
6		1	0	1	0	1	3	1.13
7		1	1	0	0	0	2	.00
8		1	1	0	0	0	2	.00
9		1	0	1	0	0	2	.44
10		1	0	0	1	0	2	.59
11		0	1	1	0	0	2	.74
12		0	1	0	1	0	2	.88
13		1	0	0	0	0	1	.00
14		1	0	0	0	0	1	.00
15		0	1	0	0	0	1	.45
16		0	0	1	0	0	1	1.14
17		0	0	0	1	0	1	1.36
18		0	0	0	1	0	1	1 36
Item Total	$n_{.j}$	12	10	7	6	3		
Sato's Caution Index	c_j	.30	.28	.42	.95	.21		

SOURCE: Harnisch & Linn, 1981

indices that have been proposed to analyze the three parts of the chart -- students (rows), items (columns), and the total test (rows and columns). Sato proposed two indices to help with an interpretation of the chart. The first index, the disparity coefficient or coefficient of heterogeneity, estimates the extent to which a test forms a Guttman scale (Guttman, 1941). When the test forms a perfect Guttman scale, the disparity coefficient equals zero. When the S-P Chart has a random pattern of item responses, the disparity coefficient is approximately one.

Sato's second index, the caution index, is a measure of the anomalousness of a response pattern, either down a column or across a row. The caution index, C for the jth item, is defined as:

$$C_j = \frac{\sum_{i=1}^{n_{.j}} (1 - u_{ij}) n_{i.} - \sum_{i=n_{.j}+1}^{I} (u_{ij} n_{i.})}{\sum_{i=1}^{n_{.j}} n_{i.} - n_{.j} \left(\frac{\sum_{i=1}^{I} n_{i.}}{I} \right)}$$

where $i=1,2,\ldots,I$, indexes the examinee,

$j=1,2,\ldots,J$, indexes the item,

$u_{ij}=1$ if examinee i answers item j correctly,

0 if examinee i answers item j incorrectly,

$n_{i.}$ =total correct for the ith examinee,

$n_{.j}$ =total number of correct responses for the jth item.

When used on the column (item), the index is intended to identify items that are not behaving properly. The index equals zero when the first $n_{.j}$ students answer the item correctly and the remaining students missed it. When the caution index equals zero, it indicates that the students with the higher scores got the item correct and the students with a lower score got the item incorrect. When the reverse pattern holds (i.e., high achievers answer incorrectly, while low achievers answer correctly), the caution index will be high. Thus, the caution index can be used in item analysis to find items that discriminate between students in an inconsistent manner. (i.e., different direction than the total scale). As can be seen from the hypothetical example in Table 1, most of the ones for items 1 and 2 are above the dashed line (representing $n._j$), so the index is low for them (.30 and .28). However, item 4 has a more random pattern resulting in a higher index (.95).

The caution index can also be used to analyze student responses. Here the index is considered to be a measure of the anomalousness of the student's response pattern. A high index indicates that the student is missing easy items and getting hard items correctly. Conversely, a low caution index indicates that the student is getting tne easier items correct and missing the hard items. The caution index can then be "used for diagnostic purposes by observing whether $C(X_i)$, the caution index, is below .5 or exceeds .5 and simultaneously noting whether the student's score is 'high' (say, above median) or 'low'" (Sato, 1980, p. 20). The double dichotomy forms four groups of students. According to Sato, students low on the caution index and high in achievement are doing well. Students low on the caution index and low in achievement need more study. However, a high caution index indicates problems other than how well the student knows the material. If the student is a high achiever, a high index means that the student is making careless mistakes. If the student is a low achiever, a high index shows that the student is not ready for the material, but might be getting a few right answers because of guessing. Again the hypothetical example in Table 1 shows students with only the easiest items correct resulting in an index equal to 0 (e.g., students 1, 3, 7, and 8). However, students having atypical responses have higher indices (e.g., students 6, 16, 17, and 18).

The caution index was used in a study by Harnisch and Linn (1981), along with a modified caution index (modified to range from 0 to 1) and the "U" index developed by van der Flier (1977). Each of the three indices were used as dependent variables to see what influenced the anomalous response. It was found that the indices were related to school (i.e., ANOVA by school). The evidence pointed toward school curricula as a reason for the response patterns, instead of the inappropriateness of a student response. In addition, Harnisch and Linn found that the three indices were highly correlated (.95) and the results obtained for the three indices were highly similar.

Patterns of Error Response

The errors that students make can also be informative about instructional and program differences. Errors can occur in two ways. Random unsystematic errors can occur for a number of reasons. A student may be guessing, or an error may have occured because of a step missed in a correct algorithm (e.g., a mistake in addition in a word problem item) that will not usually happen. Random unsystematic errors are of no use to a diagnostician except to say whether a problem is right or wrong.

A second kind of error is systematic. A student's errors are systematic when there exists an algorithm or procedure which will produce the same erroneous response over a number of similar problems. That is, given multiple problems of the same type, the student will follow the same procedure (Glaser, 1981). The diagnostician can then effectively diagnose the error behavior and help the student correct it.

Error analysis has been used to diagnose student responses in the classroom for some time. However, error analysis has been used in the research context only since the late 1970s (Birenbaum, 1980; Birenbaum & Tatsuoka, 1980; Brown & Burton, 1978; Burton, 1981; Glaser, 1981; Tatsuoka, Birenbaum, Tatsuoka, & Baillie, 1980).

Error analysis has been found to play an important role in item response theory (Lord, 1980; Warm, 1978). Tatsuoka and associates (Birenbaum & Tatsuoka, 1980; Tatsuoka, et al., 1980; Tatsuoka & Tatsuoka, 1980) found that aberrant response patterns can have an effect on the dimensionality of the data. Students with a systematic error response do not always answer problems incorrectly. Instead, an incorrect algorithm can lead to an incorrect response for some problems, and a correct response, using the same incorrect algorithm, may result for other problems. For example, students may develop different algorithms for adding a positive number with a negative number. One incorrect algorithm would be to take the difference between the two numbers and take the sign from the first number. Thus, an incorrect response will result when the first number is the smaller number (e.g.,

(-6)+9=-3 or 6+(-9)=3). However, when the first number is larger than the second number, an incorrect algorithm leads to the correct answer (e.g., (-9)+6=-3 or 9+(-6)=3). Consequently, students can answer some problems correctly, while using an incorrect method to arrive at the answer. This phenomenon has been found to affect the dimensionality of the data. While item response theory assumes unidimensionality, the assumption has been violated when an incorrect algorithm leads to a correct response. However, Tatsuoka and associates found that marking an item incorrect when an incorrect algorithm led to a correct response preserved the unidimensionality of the data.

Brown and Burton (1978) also found that different algorithms can lead to the same incorrect response. Because two different incorrect algorithms can lead to the same incorrect response, the problem of diagnosis is further complicated. For example, the error 17 + 5 = 13 could be explained by two different algorithms. The student may be adding the carry back into the same column (i.e., 7 + 5 = 2 carry 1 = 3) or the student may be adding all the numbers disregarding column (i.e., 17 + 5 = 1 + 7 + 5 = 13). Because two algorithms can lead to the same errors, the diagnosticiasn cannot help the student until one algorithm is eliminated as a possibility. This phenomenon led Brown and Burton to develop an interactive computer diagnostic system BUGGY, and later DEBUGGY, that can be used to diagnose student errors in addition and subtraction. Also, this system has been used to train teachers in the diagnosis of errors. These diagnostic systems have been used successfully to diagnose student errors on thousands of subjects (Brown & Burton, 1978; Burton, 1981; VanLehn & Friend, 1980).

Prior research on error analysis has focused on the individual. However, errors can be conceived of as a group pheno·enon. When many of the students within the same group are making the same error, it may be a result of the group instruction. Either students could be misunderstanding the instruction or the common experiences of the students may lead to an incorrect algorithm. In either case, diagnosing a student error common to the group can be a useful tool for correcting or changing the instructional program. In the next chapter some techniques for examining error response and their relationship to instructional group will be considered.

CHAPTER 3

METHODS

This chapter will be divided into three sections. In the first section, the empirical techniques that will be employed -- both test construction and the analysis of patterns of item response -- are outlined. In the second section the data base used in the empirical examples will be described. Finally, the procedures for applying the empirical techniques to the data base will be discussed.

Analytic Strategies

In attempting to better measure the achievement differences between classrooms, two analytic strategies will be considered. The first strategy, which we call the group-level test construction, uses the empirical properties of items to build scales which are more sensitive to instructional and other between-group differences. The second strategy, to be called the analysis of patterns of item response, uses the patterns of item responses (correct or error responses) at the group level to make more definitive statements about the likely differences in instruction between groups.

Test Construction

Given an interest in measuring differences between groups, indices of item discrimination at the group level need to be considered in building indicators which are more sensitive to group-to-group differences. Five indices to be considered are:

(1) intraclass correlation;

(2) intraclass correlation - between-group correlation;

(3) between-group item-total correlation;

(4) discriminant analysis; and

(5) between-group item-instructional variable correlation.

The first two indices were discussed in chapter 2, and in Lewy (1973) and Airasian and Madaus (1976). After briefly reviewing the first two techniques, the logic behind the other three criteria for item selection will be discussed below.

Intraclass Correlation. In order to build a scale which is sensitive to group differences, it is assumed that the individual items of a scale should also be sensitive to group differences (Lewy, 1973). Thus, one appropriate index might be a measure of how well an item discriminates between groups. The intraclass correlation is an index of the proportion of variation that can be attributed to group differences. So building a scale which is sens:⁺ive to group differences might be accomplished by summing items that are sensitive to group differences (i.e., a high intraclass correlation).

Intraclass Correlation - Between-Group Correlation. While the intraclass correlation is a useful index of how well an item discriminates between groups, summing items which discriminate between groups in opposite directions may result in a scale which does not discriminate between groups. Consequently, Airasian and Madaus (1976) suggested a two-step procedure, where the intraclass correlation was used to eliminate items that did not effectively discriminate between groups. Next, the correlations between the items based on their group means are used to guarantee that items are discriminating between groups in a consistent manner.

Between-Group Item-Total Correlation. Using the intraclass correlations and the between-group item correlations will create a scale that is potentially internally consistent for measuring differences between groups. However, this procedure can rapidly become unwieldy, since there are $N(N-1)/2$ intercorrelations between N items. Because of this, a variation of a procedure that has been used to build internally consistent scales for measuring individual differences might also be applied to build scales that are internally consistent in measuring differences between groups. To build internally consistent scales for measuring differences at the individual level, the point-biserial correlation of the total test score with the individual item is often used. A logical extension would be to use the correlations of group means on the total test with the group means on the items to build a reliable scale for measuring group differences. Logically, this would result in a scale similar to the scales in the prior technique (intraclass correlation - between-group correlation), but this technique seems more appealing since both the item variances and covariances play a role.

Discriminant Analysis. Since the intent of these techniques is to choose items that discriminate between groups, another approach that might be used is a discriminant analysis, where the items at the individual level are given weights to discriminate between groups. Selecting items with a high weight on the first discriminant function would yield a scale that maximizes the differences between groups along some single dimension. This would be of interest if the dimension along which the differences exist could be defined by instructional differences.

Between-Group Item-Instructional Variable Correlation. A final approach to item selection would be to use the relationship of the items to some external variable for item selection. For example, the Beginning Teacher Evaluation Study (BTES) had some success in developing scales sensitive to instructional differences between individuals (BTES: Filby & Dishaw, 1975, 1976). However, in the BTES study, all instructional variables were measured at the student level (e.g., allocated time). Because this is not always possible due to practical situations (e.g., the time and expense that would be needed in a larger study), as well as the fact that many instructional variables cannot be measured at the student level (e.g., number of aides or money invested), the criteria used in item selection might be group-level measures (e.g., instructional materials or opportunity to learn) or ~~even~~ aggregate measures of individual-level variables (e.g., time allocations). Even when the individual-level measures of the instructional variables (e.g., instructional time) are available for the item tryout, the relationship of the items to the aggregate measure might be used for item selection, if the unit of analysis is the aggregate (class, school, or program) in the final study.

Analysis of Patterns of Item Response

Given a test that has been administered to a sample of classes or schools, methods of analyzing item data can be used to gain information about instructional and program differences. These methods fall into two categories: analyzing patterns of item response for instructional differences, and analyzing error responses to identify problems common to members of a group.

Patterns of Item Response. The caution index developed by Sato has been used at the individual level as a diagnostic tool along with the total test score. However, the anomalousness of a student response may indicate more than an individual problem. A student response pattern may be related to instructional and program differences between groups. The caution index may reflect differences in emphasis and coverage. The relationship of the caution index to group membership and instructional practices needs to be further explored.

While the caution index is a useful index of the pattern of student responses, other indices of the anomalousness of a response pattern could be just as useful. Two indices in the literature are the modified caution index (Harnisch & Linn, 1981) and the "U" index (van der Flier, 1977). However, evidence indicates that the three indices are very similar in their relationship to other variables and are highly intercorrelated (Harnisch & Linn, 1981). As a consequence, only one index will be examined. The caution index was selected because of its wide use as a diagnostic tool, albeit in Japan (Sato, 1980).

Error response patterns. When the mean achievement level is low or when the group mean on a measure of the correct response pattern is high (e.g., high group mean on the caution index), a number of explanations might be put forth, including a lack of coverage. However, an alternative explanation is that the students are learning or have arrived at an incorrect algorithm for solving the problems. When the incorrect algorithm is common to students with common experiences, it could be argued that there is something in their common experience which is influencing the incorrect response pattern.

It would be useful to be able to identify classes with a common error response pattern. The first step would be to define those groups which would potentially have a problem. One criterion might be to select those groups which were low in achievement (e.g., bottom quartile) or high in the anomalousness of correct responses (e.g., average caution index higher than .5). The low achieving groups would be examined in the hopes of finding an explanation of the low performance other t. .n no knowledge of the material. The groups who were high on an anomalousness index would be examined, since the unusual pattern may indicate some logical error which is causing the odd response pattern. Errors could be classified for the algorithm used across problems and individual classes could be examined to find errors that occur across similar problems for many of the students (e.g., at least half of the group).

Besides examining individual groups that may be having problems, the distractors on a test could be examined to find out if they are related to group membership. One measure of this would be to calculate the percent of variation in the distractors that is due to differences between classrooms. The intraclass correlations of the item distractors and for "no response" could be compared to a baseline, such as the intraclass correlations on the correct item responses.

Data Base

The data to be used in subsequent empirical analyses were taken from the Beginning Teacher Evaluation Study

(BTES: Fisher, Filby, Marliave, Cahen, Dishaw, Moore, & Berliner, 1978), which was sponsored by the California Commission for Teacher Preparation and Licensing with funds from the National Institute of Education. The study was conducted to identify effective teaching behaviors that affect student learning. In particular, the study was to explore the relationship between instructional variables and reading and mathematics achievement in grades 2 and 5.

Though the study was conducted in three phases, only the third phase is relevant to the present investigation. Phase III (1974-78) of the study was conducted by the Far West Laboratory for Research and Development (FWL). Phase III was separated into three different stages. The first stage (Phase III-A, 1974-75) was spent developing additional hypotheses on teacher effectiveness. Teachers who were determined to be extremely effective or extremely ineffective in producing student achievement were observed and interviewed in a series of special studies.

After developing a model for student learning, the FWL second stage (Phase III-A Continuation, 1975-76) was to develop and refine instruments for collecting classroom process information measured in terms of time (e.g., allocated time, engaged time). In addition, achievement tests were further developed and tested to identify items and scales which were reactive to instruction. Finally, the FWL model developed in Phase III-A was tested during Phase III-B (1976-78).

The FWL model was based on the concept of "Academic Learning Time". Academic learning time (ALT) is intended to be a measure of ongoing student learning in terms of observable classroom behavior. It is defined as "the amount of time a student spends engaged on a task that produces few student errors and which is directly related to a defined content area" (Fisher et al., 1978, p. 1-7). FWL researchers hypothesized that ongoing student learning can be measured in the ALT metric and this measure provides a new and better way of measuring effective instruction. Operationalized, the model predicts a significant relationship between student achievement and measures of ALT.

Design of the Study (Phase III-B)

During Phase III-B, achievement testing was done on four occasions: (A) October 1976, (B) December 1976, (C) April 1977, and (D) September 1977. In the six weeks between testing A and B, and the seventeen weeks between B and C, extensive data on instructional process were collected. The instructional process data came from two sources: teacher logs and observations by trained field workers. The teacher logs provided daily estimates of the time allocated to individual students within specific content areas (e.g., fractions, multiplication). The second source of instructional process data was obtained through direct observation. Trained field workers observed the target students one day per week in the A-B and B-C inter-test periods. Three sources of data were collected on individual students -- allocated time in specific content areas, engagement rates, and error rates.

Information on teacher processes were also recorded. At the student level interactive teaching processes, such as presentation, monitoring, and feedback, were observed. Finally, interviews, ratings, and self-report measures provided information on diagnosis and prescription as well as teacher aptitude and classroom environment.

Sample

The original Phase III-B sample consisted of about 50 fifth grade and 50 second grade teachers who volunteered for the study in the San Francisco Bay Area (see Howell & Rice, 1977 for sampling procedures). To minimize floor and ceiling effects and to allow for data collection at the student level, 3 boys and 3 girls were selected from each class. Selection was based on a battery of reading and mathematics subscales administered in September, 1976.

Students were selected who fell between the 30th and 60th percentile of the overall distribution. FWL used this restriction to better insure that students were "typical" second and fifth graders doing second and fifth grade work.

After selection, 28 second grade classes and 30 fifth grade classes met the criteria of having 3 boys and 3 girls in the range defined above. Some teachers dropped from the study after reconsidering their commitment, leaving 25 and 22, second and fifth grade classes, respectively. Finally, one fifth grade class was dropped for failure to keep teacher logs in the A-B period.

The BTES staff considered the remaining teachers to be a representative sample. As expected, there were more female than male teachers; the sample was ethnically mixed, varied considerably in age and years of teaching experience, and represented a considerable range of teaching style and ability.

In addition, the target student sample was similar to the non-targeted sample in sex ratio, ethnic mix, and socioeconomic status. The targeted students were approximately evenly divided on sex, were ethnically mixed, and had approximately the same distribution of socioeconomic status as the population from which they were drawn. The socioeconomic status of the targeted and non-targeted were compared by the percentages of students whose parents' occupations fell into four categories -- (A) executives, professionals, managers, (B) semi-professionals, clerical, sales workers, technicians, (C) skilled and semi-skilled employees, and (D) unskilled employees.

Achievement Measures

A battery of reading and mathematics subscales were developed by the BTES staff to be reactive to instruction. The battery of exams were 180 minutes long in both grades 2 and 5. The exams were administered in two 45 minute sessions on two different days. The second grade battery consisted of 13 subtests in reading and 12 subtests in mathematics. The fifth grade battery had 11 reading subtests and 14 mathematics subtests.

Because of the large amounts of data involved in item analysis to be described later, attention will be restricted in this study to the fifth-grade fractions subtest. Fractions was a subject area in which a great deal of instructional time and effort was expended in many fifth grade classrooms. In addition, fractions was usually not taught until December. Hence, fractions was a new subject to many fifth graders and was potentially less influenced by home background. The lack of fractions instruction prior to December also meant the subtest was not administered on occasion A (October 1976).

The fraction subtest data consisted of fifteen items administered on three occasions. The skills tested included fraction addition, fraction subtraction, reducing fractions, and finding the missing numerator or denominator in fraction equations. The items from the fraction subtest are reported in Appendix A.

Pilot Data

In addition to the BTES final study (Phase III-B), the BTES pilot data (Phase III-A Continuation) will be used for the test construction phase of this dissertation. Because of an interest in instructional variables by the BTES staff, special efforts were made to develop instructionally sensitive measures (Filby & Dishaw, 1975, 1976). Two criteria were used to enhance the likelihood that the tests would be instructionally sensitive. First, item content was checked to be sure that instructional content and test content overlapped. Next, items were checked to see if gains in achievement were related to gains in instruction (Carver, 1974). This second criterion involved testing two assumptions. First, students would perform better after instruction than before instruction. Second, students who receive more instruction would achieve higher than students who received less instruction. Consequently, the pilot study, conducted in April 1975, included test item data on a broader set of fractions items (see Appendix B) and a measure of allocated time in each content area. The sample consisted of 72 subjects drawn from 5 classrooms. Achievement tests were administered on three occasions: (A) October, 1975, (B) December, 1975, and (C) April, 1976.

Measures of ALT in the Final Study (Phase III-B)

Two sources of information were available on academic learning time --the teacher logs and direct observation. The teacher logs recorded the number of minutes allocated to each content area on each day of the A-B and B-C period for each target student (Dishaw, 1977). Then, the minutes were summed within the two time periods and prorated for any missing entry. Consequently, the teacher logs gave a single measure of allocated time per content category for each student in each of the A-B and B-C periods.

The direct observation also provided an estimate of allocated time (Filby & Marliave, 1977; Fisher, Filby, & Marliave, 1977), as well as engagement rate and error rates. There were two observers per class who each rotated between 8 different classes. Observers were to code events during reading and mathematics instruction. The events were coded along three dimensions: content category, error rate ("low", "medium", or "high"), and engagement (engaged or not). An event for each targeted student was recorded once every four minutes for a full day once a week. Finally, events were summed over the A-B and B-C periods. Inter-observer reliability (Winer, 1962) was also recorded. For fractions the estimates of inter-observer reliability were .95 and .98 for the A-B and B-C periods, respectively.

Comparison of the teacher logs and observations showed a high correlation for allocated time within content area (.90 and .91 for the A-B and B-C periods, respectively). However, teacher logs were used in the final analyses. To insure consistency in the ways that allocated time was reported by teachers, an adjustment coefficient, based on the congruency of teacher logs and observer logs on the days the observers were in the classroom, was used (Marliave, Fisher, & Dishaw, 1977). Finally, engagement rates and student error rates are the ratio of the total engaged, "high" error, or "low" error time observed over the total allocated time. Student success rates were recorded as "low", "medium", and "high" (see Marliave, Fisher, & Dishaw, 1977). The low level was recorded when the student was able to perform a task with no errors except those attributed to chance (carelessness). The high level was recorded when the student was not able to respond correctly except by chance (guessing). All other activity was recorded as medium level. Both the error rates and the engagement rate were recorded without regard to specific content.

Data Analysis

Each of the analytic techniques described in the first section of this chapter will be examined using the BTES data as an empirical example. The test construction techniques will use both the final study (Phase III-B) and the pilot data (Phase III-A Continuation). The analysis of patterns of item response will focus exclusively on Phase III-B. Only the fifth grade fractions test will be used for both sets of techniques.

Test Construction

The five techniques listed in section one of this chapter will be used to construct scales. In this analysis, items will be selected on the basis of their empirical characteristics in the final data (selection criteria are described below).

Once scales are formed, two criteria will be used to examine the techniques. The intraclass correlation of the new scale will be compared to the intraclass correlation of the total scale. A difference of .05 between the two coefficients represents an increase or decrease of five percent of the between-class variation relative to the total student level variation and is assumed to represent more than a chance occurrence.

Second, the relationship of the newly formed scale to the ALT variables in the final study will be examined. This is accomplished by regressing the new scale after instruction (i.e., occasion C) on the ALT variables (i.e., allocated time, high error rate, low error rate, and engagement rate) and a pretest (i.e., the same scale prior to instruction on occasion B). The regression will be done using a contextual effects model (Alwin, 1976; Boyd & Iverson, 1979). That is, the independent variables are entered at both the group level (i.e., class means) and the individual level (i.e., student scores). The dependent variable is at the individual level. This gives an estimate of the group effect after controlling for individual differences and an estimate of the individual effects after controlling for group differences (i.e., the between-group and within-group effects). Again, the same regression using the total scale will be used as a baseline for comparisons. Effects will be considered to be substantively different when there is a difference of .05 or more in the standardized regression coefficients. The standardized regression coefficients are used because of the differences in scale length.

One final criteria for examining the technique will be to consider the stability of the empirical characteristics used in item selection. That is, would the same items be selected in the two different samples (i.e., comparing the indices in the final study and the pilot study)?

The two criteria (scale intraclass correlations and regression models) used for examining the new scales are the same across all five techniques. Therefore, they will not be further elaborated. However, the methods of item selection and of measuring stability differ from technique to technique. These will be discussed below for each technique.

(1) Intraclass Correlation. Items will be selected on the basis of a rank order of the coefficients. Scales will be formed from the five and ten items with the highest intraclass correlations. In addition, a scale will be formed of all items that reflect ten percent or more between-group variation. That is, one scale will contain all items with an intraclass correlation greater

than or equal to .10.

The stability of the intraclass correlation could be measured by a Pearson product moment correlation if more than fifteen items were involved. However, a correlation based on only fifteen pairs of numbers (i.e., index in the two samples paired by the same item) does not have enough power to be worthwhile. Thus, a comparison of the items forming the scales in the two samples will be made, taking into account the probability of randomly selecting N common items from the two samples. That is, forming a ten-item scale twice from the same 15 items, what is the probability of randomly selecting 7 (or 8 or 9) of the same items in the two samples?

(2) Intraclass Correlation - Between-Class Correlation. Airasian and Madaus (1976) used the between-group inter-item correlation in conjunction with the intraclass correlation in order to ensure that the items were discriminating between groups in the same direction. Again the approach used by Airasian and Madaus will be used. Airasian and Madaus used the two criteria in a stepwise procedure. First, items were selected on the basis of some cut score for the intraclass correlation. Next, the remaining pool of items were grouped from an examination of the between-group item correlation matrix.

As Airasian and Madaus did, the procedure will be operationalized in two steps. Using a single step would practically eliminate the use of the intraclass correlation, since on an N item test, there are N pieces of information from the item intraclass correlations and N(N-1)/2 pieces of information from the item between-class correlation matrix. Consequently, items will be used only if their intraclass correlation is greater than or equal to .10. That is, only when 10 percent or more of the variation in item performance is between groups will an item be used in the second step. The second step will be to form five-item and ten-item scales from the average inter-item correlations. As with the intraclass correlations, the stability of the index is assessed by comparing scales formed from the pilot data with scales formed from the final data.

(3) Between-Group Item-Total Correlation. This selection technique and the testing of the stability of the index will be similar to those outlined for the intraclass correlation technique. Multiple cutoffs will be used. The magnitude of the between-group item-total correlations should run high, so cutoffs will be set at .6,.7,.8, and .9. Also, as with the intraclass correlation technique, the stability of the index will be

tested by comparing scales formed in the pilot st.`y with the scales formed in the final study using the same index.

(4) Discriminant Analysis. The discriminant function presents a problem because the discriminating variables are assumed to be normal. Instead, they are binary items. In the last decade, analysis of binary data has received a great deal of attention from methodologists. The advances in the factor analysis of dichotomous data (Muthen, 1978, 1980; Muthen & Christoffersson, 1979) and the method of logistic discrimination (Anderson, 1974, 1979; Cox, 1966; Day & Kerridge, 1967) are evidence of these developments. However, without the availability of a computer package to solve the maximum likelihood iterative procedure, a discriminant function, assuming normally distributed discriminating variables, will be used, while recognizing the possible bias in the procedure. The stability of the technique will be assessed by comparing the number of functions derived in the two samples and any similarities in the standardized canonical discriminant coefficients.

(5) Between-Group Item-Instructional Variable Correlation. The instructional variable used in this analysis is allocated time (the only variable available in the pilot study). Similar to the intraclass correlation and the between-group item-total correlation techniques, cutoffs will be set for a single index. The cutoffs will

be used to form five and ten item scales with the highest correlations. Also, similar to the analyses for techniques (1), (2), and (3), the stability of the index will be tested by comparing the five-item and ten-item scales formed in the two different samples.

Analyzing and Interpreting Patterns of Item Response

The analysis of item data to gain information about instructional differences across classrooms will be applied to data from the final study (Phase III-B). Students were tested on three occasions. The occasions were prior to instruction, after instruction, and after a summer break. These three different points in the instructional sequence of fifth grade fractions leads us to expect certain patterns of results from the data in the presence of effective instruction. For example, the mean performance on items should increase with instruction. In addition, a slight decrease in item performance should occur over the summer break (e.g., forgetting or confusing algorithms for different problems). Thus, one would expect the highest performance on occasion C, with occasion D being greater than or equal to occasion B performance. (Summer loss should have a floor effect defined by knowledge prior to instruction.) The instructional sequence and the relevant timing of the achievement tests will be used to explain differences in patterns of item response from one testing to another.

(1) Student Response Patterns. Sato's caution index (Sato, 1980) will be used to examine the anomalousness of student response patterns. If the pattern of responses is related to instruction as is hypothesized, the caution index should show more between-class variation after instruction than before instruction. In addition, the experiences (i.e., forgetting or learning through practical experiences such as using money) that influence achievement over the summer would not be class related and consequently, the between-class variation would decrease. Thus, the same pattern would be expected of the intraclass correlation of the caution index as of the mean performance.

Besides the behavior of the caution index over time, some hypotheses might be made about its behavior after instruction. First, those classes with a mean on the caution index above .50 should have a different pattern of responses than the total sample. The class mean on an index of anomalousness will be affected by guessing and carelessness, but the highest mean would be expected when all the students in a class are uniformly high. In the presence of a different (from the total sample) pattern of coverage and emphasis, students should have a uniformly high caution index. Thus, a high class mean on the caution index may indicate an instructional group with a peculiar pattern of coverage and emphasis.

Second, the relationship of the caution index to the instructional variables will be explored. This is accomplished by using the same contextual effects model used in the test construction phase of this dissertation, except with the caution index used as the dependent variable (i.e., student anomalousness as function of a pretest and the ALT variables). Finally, the possibility that the caution index differs between classes in a way not related to the ALT variables or the achievement test will be examined. This is accomplished by running a simple ANOVA with the classes used as the groups.

(2) Error Response Patterns. Incorrect responses will be analyzed in two ways. First, the distractors will be examined to find out which distractors are the most influenced by class structure. Second, individual classes will be examined to find out what errors are common throughout the class.

As with the caution index, the intraclass correlation of the distractors should be related to test occasion in the presence of instructional effects. Hence, the intraclass correlations should increase on occasion C and decrease on occasion D.

It also seems worthwhile to examine the problems occurring in individual classes. This examination will consist of finding those error responses which occur across similar problems for half or more of the students in a class. That is, an error is considered to be common to a class when half or more of the students in the class select the same incorrect response across similar problems.

CHAPTER 4

SUBSETS OF GROUP SENSITIVE ITEMS

In this chapter the possibility that classroom differences can be better defined by subgroups of items within a subtest (fractions) is explored. This chapter is divided into two parts. The major thrust of the chapter is to determine if subsets of items in Phase III-B of the BTES study can be identified, which are more sensitive to classroom differences than the total score. Thus, the first section of this chapter will describe the data from Phase III-B of the BTES study that will be used in this analysis (i.e., the fifteen item fractions subtest and the measures of ALT).

Besides selecting subsets of group sensitive items from the final study data (Phase III-B), the stability of the item selection indices is explored by comparing the indices in the final study with the analogous indices in the pilot study (Phase III-A Continuation). So the first section of this chapter (descriptive statistics) will also describe the data needed for this analysis in Phase III-A Continuation of the BTES study. Then the second section of this chapter will be the analyses proposed in Chapter 3 for analyzing the empirical procedures for forming subsets of group sensitive items.

Descriptive Statistics

Final Study (Phase III-B)

The fifteen-item fractions subtest (see Appendix A for a copy of the test) was administered on three occasions. The three occasions might be defined by the typical fifth grade instructional agenda in fractions. The first testing (occasion B) was prior to instruction in the area of fractions. The second testing (occasion C) was after instruction. Finally, the students were tested after a summer break (occasion D).

Item Means. The relationship of the test administrations to the instructional calendar leads to certain notions which will be examined. For example, achievement should be higher after instruction than prior to instruction. In addition, while the summer should result in a small loss in achievement (students will not retain some skills which are not exercised during the break), this loss should have a floor effect defined by achievement prior to instruction. That is, the students will not lose more knowledge than they had gained.

The means for the fifteen items on the three occasions are reported in Table 2. As expected, the means on most of the items and on the total test rose sharply

Table 2. Phase III-B fraction item means.

	Occasion		
ITEM	B	C	D
1	.49	.79	.69
2	.45	.76	.72
3	.47	.53	.53
4	.40	.66	.62
5	.11	.35	.19
6	.43	.69	.55
7	.34	.62	.47
8	.11	.26	.30
9	.07	.24	.28
10	.32	.54	.53
11	.32	.55	.46
12	.22	.47	.51
13	.15	.32	.28
14	.29	.53	.51
15	.20	.33	.40
N	127	123	89
Total Test	4.37	7.64	7.04

after instruction, and fell slightly after the summer break. To check for any possible bias from attrition, the item means on occasions B and C were calculated using only the 89 cases present on occasion D.[1] Comparing these means to the means in Table 2, it was concluded that no attrition bias was present.

It should also be noted that three of the most difficult items are those that require multiple skills (i.e., 5, 8, and 9). While most items require addition, subtraction, or recognizing equivalences, these three items require either addition and equivalent forms of a fraction (i.e., 5) or subtraction and equivalent forms of a fraction (i.e., 8 and 9). These three items are expected to be more difficult than most of the remaining items since multiple skills gives the student multiple ways to get the wrong answer.

Intraclass Correlations. As with the item means, the trend in the item intraclass correlations across the three occasions can be logically explained. Instructional differences between classes should help to strengthen the relationship of achievement to class membership. Thus, one would expect an increase in the intraclass correlation after instruction and a decrease over the summer break to some value not lower than prior to instruction.

The item intraclass correlations are reported in Table 3. As can be seen, the mean value on occasion B (.26) shows that prior to instruction, the grouping mechanism is related to achievement. While students can be randomly assigned to classrooms within a school, the students at the same school are usually very similar in socioeconomic status and prior educational experience. Thus, the common background of students living in the same neighborhood, attending the same school, and assigned to the same classroom accounts for approximately 26 percent of the variation in item achievement prior to instruction.

A common curricula serves to strengthen the relationship between class membership and achievement. Thus, the mean intraclass correlation was .07 higher after students within each class shared a common instructional program (.33). Finally, the summer break dampens the effect of a common curricula. In fact, the mean value after the summer break is the same as prior to instruction (.26). So, the summer results in an increase in the within-class variation.

While the intraclass correlations for the items increase with instruction, the same is not true for the total test. The intraclass correlation decreases on each occasion, indicating that the within-class variation is increasing faster than the between-class variation with or

Table 3. Phase III-B fraction item intraclass correlations.

		Occasion	
ITEM	B	C	D
1	.38	.39	.21
2	.36	.38	.23
3	.17	.25	.10
4	.26	.27	.36
5	.26	.31	.33
6	.26	.27	.28
7	.38	.33	.33
8	.15	.46	.34
9	.20	.36	.26
10	.25	.39	.34
11	.26	.27	.21
12	.29	.37	.23
13	.22	.26	.23
14	.24	.39	.22
15	.26	.28	.26
Mean	.26	.33	.26
Total Test	.50	.47	.42

without instruction. This phenomenon, in conjunction with the increase in the mean item intraclass correlation, may be a result of differential coverage of the materials in the test. That is, teachers cover different subsets of items. Thus, for any given item there is differential instructional coverage across classrooms leading to high intraclass correlations. But, teachers who cover one item may not cover another, so that some inter-item covariances are negative. Since the total score variance contains both item variances and item covariances, the differential topic coverage leads to negative covariances and thus reduced total test intraclass correlations.

Item Intercorrelations. The item correlations at the between-class and within-class level are reported in Tables 4, 5, and 6 for occasions B, C, and D, respectively. Examining the three between-class correlation matrices supports the notion of differential coverage of materials for classes. For example, items 14 and 15 (the only algebraic manipulations items) have high positive between-class correlations with all items on occasions B (pretest) and D (after summer break). However, instruction causes the between-class correlations of items 14 and 15 with items 1 to 10 to be greatly reduced and in some cases negative. Thus, without

Table 4. Phase III-B item intercorrelations between classes (lower triangle) and within class (upper triangle) on occasion B.

	Subtraction					Addition					Equating			Algebraic Manipulation	
ITEM	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		.72	.00	.36	.05	.31	.26	.26	.11	.14	.24	.11	.03	.16	-.03
2	.85		.15	.33	-.07	.38	.40	.19	.18	.21	.15	.02	-.11	.10	-.05
3	.39	.27		.28	-.06	.16	.21	.19	-.00	.08	.12	-.09	.03	-.02	-.12
4	.67	.76	.38		-.16	.21	.24	.13	.05	.21	.26	.16	.03	.16	.10
5	.08	.10	.60	.13		-.00	-.08	.17	.00	.06	.01	-.08	.05	-.07	.05
6	.70	.68	.46	.44	.21		.51	.23	.15	.28	.22	.02	.06	.05	-.02
7	.72	.74	.47	.64	.22	.91		.16	.13	.27	.19	.20	.08	.08	.10
8	.50	.64	.42	.59	-.08	.63	.66		.32	.26	.14	.14	.07	.07	.20
9	.31	.44	.29	.50	.00	.62	.60	.73		.15	.23	.03	-.08	-.05	.08
10	.46	.60	.49	.55	.15	.59	.57	.71	.65		.29	.15	.04	.13	.14
11	.18	.23	.59	.31	.31	.43	.44	.42	.42	.60		.27	.30	.12	.25
12.	.50	.61	.50	.40	.20	.80	.76	.62	.54	.64	.66		.27	.06	.32
13	.35	.36	.39	.35	.10	.54	.54	.44	.57	.26	.43	.65		.01	.13
14	.54	.64	.56	.51	.41	.74	.73	.56	.59	.53	.45	.75	.69		.43
15	.54	.62.	.52	.74	.15	.59	.66	.56	.49	.59	.62	.64	.57	.76	

Table 5. Phase III-B 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	[illegible]	.22	.26	.20
2	.86		.00	.29	-.08	.54	.33	.07	-.06	.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	

Table 6. Phase III-B item intercorrelations between classes (lower triangle) and within classes (upper triangle) on occasion D.

	Subtraction					Addition					Equating			Algebraic Manipulation	
ITEM	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		.86	.18	.46	.13	.35	.50	.33	.25	.26	.19	.16	-.02	.09	.08
2	.89		.08	.36	.12	.38	.46	.30	.23	.26	.14	.18	-.08	.15	.15
3	.55	.61		.33	.02	.22	.20	.23	.11	.33	.18	.09	.08	.19	.09
4	.49	.62	.62		-.09	.22	.31	.07	.19	.25	.04	.13	-.08	.02	.06
5	.66	.57	.55	.42		.38	.37	.42	.29	.32	.09	.07	.05	.21	.21
6	.59	.62	.05	.22	.58		.68	.36	.44	.52	-.03	-.07	-.02	.01	.04
7	.69	.76	.34	.49	.64	.82		.45	.49	.52	.02	.22	.02	.18	.21
8	.78	.72	.56	.64	.82	.65	.75		.29	.45	.20	.21	.19	.09	.19
9	.77	.67	.32	.51	.66	.71	.75	.86		.43	.17	.16	.09	.22	.14
10	.72	.81	.36	.60	.56	.77	.83	.83	.82		.08	.07	.02	.18	.22
11	.65	.52	.31	-.00	.64	.53	.39	.47	.40	.34		.33	.26	.30	.29
12	.79	.63	.50	.37	.74	.41	.52	.76	.62	.51	.52		.34	.48	.38
13	.52	.52	.14	.12	.60	.69	.69	.57	.64	.63	.54	.46		.33	.24
14	.62	.54	.51	.47	.58	.21	.48	.55	.43	.40	.21	.80	.27		.61
15	.44	.39	.38	.45	.70	.25	.42	.58	.56	.42	.11	.67	.27	.70	

instruction, classes high on one item are high on another, but the immediate effect of instruction is to decrease the relationship of items from different content areas through differential coverage and emphasis.

In addition to instruction diminishing the between-class relationship of items with different contents, the between-class correlations of similar items appears to be strengthened. That is, clusters of items similar in content can be picked out using the between-class correlations on occasion C, while the common content patterns of correlation are not as obvious with the within-class correlations. For example, the between-class correlations of the five addition items (i.e., items 6 through 10) range from .53 to .86. In contrast, the range of between-class correlations of the five items with the other ten items is -.12 to .69. This pattern is not as obvious with the within-class correlation matrix, where the correlations between the same five items ranges from .14 to .67 and the correlations with the other ten items range from -.00 to .54. Thus, the overlap between the two sets of within-class correlations (i.e., items 6 through 10 with themselves and with the other 10 items) is larger than the overlap of the same two sets of between-class correlations. Also, the within-class correlations for the five items are much lower in magnitude than the same correlations between-class (.14 as opposed to .53).

There are two possible explanations for why the relationship between classes is stronger than the relationship within classes. The first is that the instruction or the prior educational experiences are effective. That is, all students in the classroom are learning the same skills because of the school or class curricula. However, another explanation might be the sampling procedures used to collect these data. By selecting students from the midrange of the distribution, class means might still be measured accurately. However, selecting subjects from the midrange will definitely reduce the variance within groups. Thus, the within-class correlation matrix might be inappropriately measuring relationships because of a reduction in the variances of the variables and the dissimilarity of the between-class and within-class correlation matrices may be an artifact of the sampling technique employed.

Total Scale Statistics. Descriptive statistics for the total scale are found in Table 7. The means were as expected with a large increase after instruction and a small summer loss. The internal consistency coefficients (Cronbach, 1951) were also high at both the individual and the class level. However, the differential coverage of

Table 7. Phase III-B total scale descriptive statistics -- individual and class level.

	Occasion		
	B	C	D
Mean	4.37	7.64	7.04
Individual			
S.D.	3.49	4.05	4.24
Reliability	.82	.86	.87
Correlations			
B			
C	.53		
D	.51	.82	
Class			
S.D.	2.47	2.76	2.77
Reliability	.94	.92	.95
Correlations			
B			
C	.67		
D	.75	.93	

materials at the class level leads to a slight decrease in the internal consistency after instruction. The correlations at both levels were also as expected, with the correlation being highest for the two occasions after instruction (occasions C and D). Finally, the standard deviations are consistent with prior educational research. At the individual level, the standard deviation increases with each testing. Even in the absence of instruction, the low achievers continue to fall behind, while the high achievers continue to move ahead. At the class level, the standard deviation increases only as a function of their common experiences. Differences in the effectiveness of an instructional program will increase the variance between groups. However, in the absence of any common experience as yet unaccounted for, the between-class variance does not increase (i.e., over the summer break).

Academic Learning Time. Finally, descriptive statistics for the academic learning time variables and their correlations with fractions achievement are reported in Table 8. The average student spends approximately 44 minutes a week (750.45 minutes between occasions B and C divided by the 17 weeks of instruction during the period) studying fractions, which is more time than is spent in any other area of mathematics in the fifth grade. Of those 44 minutes, the average student spends 33 minutes

Table 8. Phase III-B instructional variables descriptive statistics and correlations with fractions subtest.

	Allocated Time (AT)[a]	Engagement Rate (ER)	High Error Rate (HER)	Low Error Rate (LER)
Mean	750.45	.75	.03	.36
S.D.	724.84	.25	.05	.18
Intraclass Correlation	.72	.60	.27	.52
Individual Level Correlations				
AT				
ER	-.11			
HER	-.14	.05		
LER	-.30	.23	-.05	
Fraction Test Occasion:				
B	.09	.12	-.10	.05
C	.41	.22	-.27	-.04
D	.32	.19	-.18	-.02
Class Level Correlations				
AT				
ER	-.08			
HER	-.14	.11		
LER	-.51	.14	.36	
Fraction Test Occasion:				
B	.13	.09	-.10	-.20
C	.61	.24	-.31	-.40
D	.47	.29	-.34	-.35

[a] For scaling purposes, allocated time is transformed to minutes per day in the analyses in this chapter. Mean allocated time per day is 750.45 divided by 17 weeks divided by 5 days per week.

engaged in some activity, approximately 1.3 minutes with a high error rate (chance-level performance) and 16 minutes with a low error rate (high performance). Furthermore, there is more variability between classes than within classes in allocated time, engagement rate, and low error rate. Only for a high error rate is the within-class variation higher than the between-class variation.

The correlations between the ALT variables show that the student who has more success experiences (i.e., low error rate) is allocated less time (r=-.30) and is engaged more often (p=.23). At the class level, less time is allocated when there are more success experiences (p=-.51). In addition, the classroom with a low error rate also has more failing experiences (p=.36).

Finally, the correlations of the ALT variables with the fractions subtest show that the higher achieving students after instruction received more allocated time (r=.41), were engaged more often (r=.22), and were less likely to answer at a chance level during the instructional sequence (r=-.27). At the class level, achievement was higher when more time was allocated (r=.61), more time was engaged in learning (r=.24), and success rates were neither high (r=-.40) nor low (r=-.31).

Pilot Data (Phase III-A Continuation)

The pilot data (i.e., Phase III-A Continuation) had achievement tests on three occasions -- A, B, and C. Test occasion A was at the beginning of the school year (October, 1975); B was in December, 1975; and C was in April, 1975. The tests were administered to all the students in five classrooms. On the basis of fractions achievement and the time allocated to fractions instruction during the A-B Period, the BTES staff decided not to administer the fractions achievement subtest on occasion A in the final study (i.e., Phase III-B). It was assumed that the low achievement on occasion B would serve as a baseline of knowledge prior to instruction.

Item Means. The item means for the thirty fractions items (see Appendix B) are contained in Table 9. From the item means, it seems that the BTES staff was justified in not testing fractions on occasion A in the final study and in treating occasion B as a baseline for achievement prior to instruction. A completely random response pattern on a 30 item multiple choice test with four alternatives per item would average 7.5 (i.e., 30/4 = 7.5). As can be seen from Table 9, the mean achievement on fractions on occasions A and B are below what would be expected by chance (60 and 6.10 on fractions on occasions

Table 9. Phase III-A Continuation item means.

		Occasion		Item in Final Study
Item	A	B	C	
1	.16	.23	.46	11
2	.06	.10	.25	13
3	.11	.14	.27	
4	.11	.15	.43	14
5	.19	.18	.50	
6	.16	.20	.43	15
7	.09	.19	.36	12
8	.04	.10	.18	
9	.16	.16	.31	
10	.10	.16	.29	
11	.20	.32	.56	6
12	.13	.25	.49	8
13	.01	.02	.09	
14	.02	.02	.11	
15	.06	.13	.30	9
16	.05	.10	.19	
17	.07	.13	.31	
18	.14	.20	.37	7
19	.10	.20	.38	10
20	.06	.11	.17	
21	.34	.39	.64	1
22	.14	.19	.25	
23	.16	.20	.27	
24	.31	.51	.69	2
25	.29	.19	.29	
26	.27	.41	.55	3
27	.32	.37	.52	5
28	.18	.16	.16	
29	.29	.42	.47	4
30	.18	.17	.28	
Total	4.50	6.10	10.57	

A and B, respectively). In addition, mean time allocated to fractions instruction during the A-B period was 16.80 minutes, or less than three minutes per week (16.80/6).

In contrast, on occasion C achievement was above the score expected from random response. The mean on the 30 item test on ocacsion C was 10.57. In addition, the time allocated to fractions instruction during the B-C period was 612.98 minutes, or 36.06 minutes per week (612.98/17). Thus, there was more than an eighteen fold increase in fractions instruction from the A-B period to the B-C period.

Finally, the achievement pattern is the same for 15 items used in the final study as for the entire 30 items. Achievement on the 15 item exam used in the final study was well below the 3.75 mean expected by chance (2.74) on occasion A. Achievement on occasion B was slightly above the random response level (4.07). Finally, the fractions acheivement was well above random response on occasion C (6.80). In addition, the mean achievement on the fifteen items used in the final study all increased from occasion A to occasion B and from occasion B to occasion C.

Since the fifteen items used in the final study did not differ significantly from the entire 30-item subtest, all further results will focus strictly on the subset of fifteen items. Since no data is available on the other 15 items in the final study, attention will be restricted to the items used in both studies.

Intraclass Correlations. An examination of the fifteen-item intraclass correlations in Table 10 further supports the notion that instruction did not affect achievement during the A-B period. The fifteen items used in the final study were typical of the pattern of intraclass correlations for the full thirty items. There was not one item on occasion A nor occasion B with an intraclass correlation greater than or equal to .10. In other words, less than 10 percent of the variation in item response on both occasions A and B could be attributed to classroom differences. In contrast, the median intraclass correlation for the items on occasion C, using either the final 15 items or the full 30 items, was .11. So over half of the items on occasion C had over 10 percent of their variation accounted for by classroom differences.

The pattern of item intraclass correlations from the pretest to the posttest in the pilot data (Table 10) were similar to the item intraclass correlations in the final study (Table 3). That is, the item intraclass correlations were higher after instruction than prior to instruction. However, the magnitude of the intraclass correlations in the final study was much higher than in

Table 10. Phase III-A Continuation item intraclass correlations.

	Occasion		
ITEM	A	B	C
1	.08	.03	.14
2	.03	.01	.11
3	.06	.04	.08
4	.03	.09	.11
5	.08	.02	.04
6	.05	.05	.03
7	.04	.02	.16
8	.05	.03	.18
9	.02	.05	.06
10	.01	.02	.18
11	.06	.03	.11
12	.07	.03	.11
13	.03	.02	.24
14	.02	.04	.05
15	.05	.01	.10
Median	.05	.03	.11

the pilot study. This phenomenon was probably due to the restriction in range in the final study. By selecting students who were neither exceptionally high nor low in achievement, the within-class sums of squares is reduced. Thus, the intraclass correlations are increased by restricting the within-class variability, when the between-class variability is held constant.

Item Intercorrelations. Apparently, the broader within-class range of achievement in Phase III-A Continuation affects the within-class item correlation matrix (see Table 11). The between-class matrix is similar to the same matrix for the final study (Table 5) in magnitude. However, the subgrouping of the items into content cohesive groups is not as well defined as in the final study. While the coefficients of the between-class correlation matrix are generally high and positive, no subgroups of items can be readily defined. With a sample of only five classrooms, there may have been no marked differences in curriculum coverage and emphasis. The restriction in the within-class variation had little or no effect on the between-class correlation matrix. In contrast, the within-class correlation matrix is different in magnitude in the pilot study than in the final study. By removing the restriction on within-class variation in achievement, the item correlations are high and positive

Table 11. Item intercorrelations between classes (lower triangle) and within classes (upper triangle) on occasion C of Phase III-A Continuation.

	Subtraction					Addition					Equating			Algebraic Manipulation	
ITEM	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		.68	.45	.55	.50	.33	.32	.23	.21	.35	.24	.10	.21	.33	.24
2	.84		.48	.52	.58	.15	.24	.07	.22	.24	.16	.20	.30	.26	.23
3	.05	.21		.45	.61	.18	.35	.15	.17	.43	.28	.35	.28	.34	.30
4	.17	.50	.57		.55	.20	.25	.15	.13	.37	.17	.16	.30	.34	.23
5	.48	.70	.80	.84		.13	.36	.17	.10	.47	.20	.35	.37	.28	.31
6	.79	.70	.41	.62	.76		.50	.63	.48	.52	.34	.06	.05	.18	.22
7	.24	.38	.94	.76	.91	.67		.43	.48	.75	.28	.29	.32	.36	.30
8	.55	.75	.63	.90	.96	.85	.83		.49	.44	.29	.11	.11	.16	.18
9	.91	.82	.44	.45	.76	.93	.62	.79		.38	.33	.07	.17	.30	.21
10	.15	.42	.80	.95	.91	.64	.92	.90	.51		.16	.32	.22	.30	.32
11	-.15	.31	.63	.41	.54	-.11	.47	.36	.04	.48		.36	.38	-.33	.39
12	.33	.62	.88	.71	.94	.53	.88	.82	.62	.83	.77		.38	.31	.47
13	.57	.88	.56	.79	.92	.67	.69	.91	.73	.76	.60	.87		.38	.28
14	.74	.90	.48	.38	.75	.57	.52	.68	.79	.42	.55	.79	.86		.29
15	.60	.90	.45	.52	.77	.48	.49	.72	.66	.51	.68	.82	.93	.95	

as in the between-class correlation matrix.

Analyzing Subsets of Items Sensitive to Group Differences

As pointed out earlier, the use of a test score within some defined content area may mask differences between groups which can be defined at the item level. However, it is equally possible that some unit of measurement which is not the total test, but some combination of items within the test can be identified that is sensitive to group differences. This newly formed scale might be formed to be more sensitive to differences between groups through the empirical properties of items. The five empirical properties described in Chapter Three -- intraclass correlation, intraclass correlation - between-group item correlation, between-group item-total correlation, discriminant analysis approach, and item-instructional variable correlation between groups -- are used to construct scales more sensitive to group differences.

Intraclass Correlation

The intraclass correlations for the items on occasion C are contained in Table 3. Forming scales of the top five and top ten items results in scales that are mixed in content (see Table 12). The five items with the highest intraclass correlations (.38 and above) include

Table 12. Fractions subtest (occasion C) regressed on ALT variables and pretest (occasion B).[a]

	REGRESSION COEFFICIENTS	
	Unstandardized	Standardized
Between Class		
Pretest	.24 (.55)	.14
Allocated Time	.14 (.67)	-.24
Engagement	1.07 (.17)	.05
Low Error	-1.91 (.20)	-.06
High Error	-8.74 (.26)	-.06
Within Class		
Pretest	.38 (3.19)	.34
Allocated Time	.08 (1.15)	.16
Engagement	2.50 (1.29)	.16
Low Error	2.44 (.94)	.11
High Error	-11.76 (1.83)	-.16
Constant	.81	
R^2	.52	

[a] t-statistics within parenthesis - between class df=15, within class df=96.

subtraction (items 1 and 2), addition (items 8 and 10), and solving the equation (item 14). The new scale also has the two simplest items (1 and 2), the next to the hardest item (8), and the two items in the middle of the difficulty distribution (10 and 14). Apparently, the scales formed by the intraci .s correlation (both the top 5 and top 10) are unrelated to item difficulty and item content.

For comparative purposes, the 15 item fractions scale is regressed on the pretest and the ALT variables in Table 13. The two most important determinants of fractions achievement are within-class pretest and between-class allocated time. Both have a positive effect. Thus, the high achiever in fractions would be in a class with a high amount of time allocated to fractions and the student would have been a higher performer in fractions relative t. his classmates at the pretest. The intraclass correlation for the fractions test on occasion C was .47, indicating the almost half of the variation in fractions achievement was between classes.

The analogous regression equations for the two new scales are contained in Table 14. Also, the correlations of these scales and all other scales formed in this chapter with the pretest and the ALT variables are contained in Appendix C. The results of these scales are

Table 13. Five and ten item scales formed from the intraclass correlations regressed on the same pretest and the ALT variables.[a]

	REGRESSION COEFFICIENTS			
	Unstandardized		Standardized	
	5 items	10 items	5 items	10 items
Intraclass Correlation	.46	.49		
Between Class				
Pretest	.26 (.67)	.20 (.48)	.17	.12
Allocated Time	.05 (.60)	.11 (.83)	.23	.30
Engagement Rate	.83 (.33)	2.31 (.55)	.11	.16
Low Error Rate	-2.47 (.65)	-2.37 (.36)	-.21	-.11
High Error Rate	3.10 (.23)	-3.78 (.16)	.05	-.04
Within Class				
Pretest	.19 (1.97)	.36 (3.10)	.20	.31
Allocated Time	.03 (1.03)	.05 (1.12)	.15	.15
Engagement Rate	1.26 (1.64)	1.36 (1.04)	.21	.12
Low Error Rate	1.21 (1.19)	2.27 (1.30)	.15	.15
High Error Rate	-5.37 (2.10)	-7.73 (1.78)	-.19	-.15
Constant	.43	-.40		
R^2	.46	.53		

[a] t-statistics within parenthesis - between class df=15, within class df=96.

Table 14. Overlap of subtests formed from the pilot study and the final study using the intraclass correlations.

	Top 5		Top 10	
Item	Pilot	Final	Pilot	Final
1	x	x	x	x
2		x	x	x
3				
4			x	
5				x
6				
7	x		x	x
8	x	x	x	x
9				x
10	x	x	x	x
11			x	
12			x	x
13	x			
14		x		x
15			x	x
Overlap (number of items)	3		7	

contrary to the findings of Airasian and Madaus (1976). That is, the scales formed using the intraclass correlation do not increase the intraclass correlation of the scale. The intraclass correlations for the two scales (.46 and .49) are in the same range as the intraclass correlation for the total scale (.47).

Next, the regression equations are compared to the same equation for the total scale (Table 13). The standardized regression coefficients are used because the different scales do not contain the same number of items and thus are not in the same metric. Most of the coefficients are comparable, but some coefficients indicate that the new scales are more sensitive to between-class differences as opposed to within-class differences. The within-class coefficients for the ALT variables and the between-class coefficients for the pretest are comparable for all the scales. However, the importance of the within-class pretest is reduced from .34 to .31 and .20 for the two scales. In addition, the effect of the ALT variables between classes is increased. Coefficients for engagement rate increase from .05 to .16 and .11. Low error rate coefficients increase in absolute magnitude from -.06 to -.11 and -.21. The trend is less clear for allocated time and high error rate.

In summary, the scales formed from the item intraclass correlations result in a scale less sensitive to differences in entering ability within class, and more sensitive to engagement rate and low error rate differences between classes. Students achieve higher when they are from a class which spends a greater proportion of their time engaged in their work, perhaps indicating a more positive attitude and a class which spends less time on lower error rate activities, perhaps indicating more challenging work.

To tast the stability of the item intraclass correlations across samples, the items used for a five and ten item scale based on the item intraclass correlations for the pilot data (see Table 10) and the final data (Table 3) are reported in Table 12. As can be seen three items were the same on the five item scale and seven items were the same on the ten item scale. To better understand this, the probabilities of having different numbers of common items for a five and ten item scale by chance are reported in Table 15. So, the probability of having seven or more items in common on a ten item scale by chance is very high (p=.57), while the probability of having three or more items in common on two five item scales is much lower (p=.17). However, the combined results of the two scales indicate that the item intraclass correlations are sample dependent (i.e., not stable across samples).

Table 15. Probability (p) of getting N_i common items randomly by selecting N items on a fifteen item exam.*

N_i (given N=5)	N_i (given N=10)	p
5	10	.00
4	9	.02
3	8	.15
2	7	.40
1	6	.35
0	5	.08

$$* \quad p = \frac{\binom{N}{N_i}\binom{15 - N}{N - N_i}}{\binom{15}{N}}, \text{ where } \binom{I}{J} = \frac{I!}{J!(I - J)!}$$

Intraclass Correlation - Between-Class Item Correlations

The intraclass correlation played no part in the scale construction because all items met the criterion of having an intraclass correlation greater than .10 (see Table 3). Thus, this technique reduced to using the average between-class item correlation (see Table 5) as reported in Table 16. Again, the top five and top ten items on the average between-class item correlation do not form a content cohesive group of items nor do they appear to be related to item difficulty (see Table 17).

As with the prior item selection strategy, this technique does not lead to a greater proportion of variation between classes in the newly formed scales than in the total scale (see Table 18). However, the scale appears to be more instructionally sensitive than the total scale. While the differences between the ten-item scale and the total scale are small, trends in the data can be seen by comparing the coefficients for all three scales (five, ten, and total) simultaneously. The importance of both allocated time and engagement rate between classes increases, while the importance of the same two variables within classes decreases. Thus, the scale is more sensitive to instructional differences

Table 16. Average between-class item intercorrelations in the final and pilot study.

	Average Between-Class Correlations	
Item	Pilot	Final
1	.45	.49
2	.64	.38
3	.56	.28
4	.61	.48
5	.79	.45
6	.61	.42
7	.67	.42
8	.70	.51
9	.65	.53
10	.66	.41
11	.40	.53
12	.74	.49
13	.77	.45
14	.67	.27
15	.68	.36

Table 17. Overlap of subtests formed from the pilot study and the final study using the average between-class item intercorrelations.

	Top 5		Top 10	
Item	Pilot	Final	Pilot	Final
1		x		X
2				
3				
4			x	x
5	x		x	x
6				x
7			x	x
8	x	x	x	x
9		x	x	x
10			x	
11		x		x
12	x	x	x	x
13	x		x	x
14			x	
15	x		x	
Overlap (number of items)	2		7	

Table 18. Scales formed from the average between-class item correlations regressed on the same pretest and ALT variables.[a]

	REGRESSION COEFFICIENTS			
	Unstandardized		Standardized	
Cutoff	Top 10	Top 5	Top 10	Top 5
Intraclass Correlation	.48	.46		
Between-Class				
Pretest	.30 (.61)	.30 (.61)	.16	.14
Allocated Time	.10 (.66)	.08 (.99)	.25	.35
Engagement Rate	1.54 (.33)	1.38 (.56)	.10	.17
Low Error Rate	-.72 (.10)	-1.08 (.29)	-.03	-.09
High Error Rate	-8.31 (.33)	-1.93 (.14)	-.08	-.03
Within-Class				
Pretest	.44 (3.34)	.50 (4.35)	.35	.40
Allocated Time	.04 (.88)	.00 (.11)	.13	.01
Engagement Rate	.62 (.43)	.14 (.18)	.05	.02
Low Error Rate	1.28 (.67)	1.41 (1.41)	.08	.16
High Error Rate	-6.79 (1.43)	-3.54 (1.33)	-.13	-.12
Constant	.41	-.48		
R^2	.48	.53		

[a] t-statistics within parenthesis - between-class df=15, within-class df=96.

between classes instead of within classes. However, the scale is still highly influenced by the within-class pretest.

To examine the stability of the average between-class item correlation, the five and ten item scales that would be formed in the pilot study and the final study (see Table 16) are presented in Table 17. Again, the probability of having only two or seven items in common on a five or ten item scale randomly is very high. From Table 15, it can be shown that the probability of randomly selecting two (or seven) or more items that are the same on two five (or ten) item tests is .57. Thus, it can be concluded that the average between-class item correlations are not stable from one sample to another in this study.

Between-Class Item-Total Correlation

Since the between-class item-total correlation was suggested as a simpler (computationally) alternative to the average between-class item correlation strategy, it is not surprising that the results are the same. In fact, an examination of the between-class item-total correlations in Table 19 shows that identical five and ten-item scales would be formed as using the average between-class item correlations, in both the pilot study and the final study. So the conclusions from this strategy are identical to

Table 19. BTES fractions between-class item-total correlations.

	Average Between-Class Correlations	
Item	Pilot	Final
1	.49	.71
2	.77	.54
3	.69	.39
4	.77	.70
5	.99	.66
6	.74	.61
7	.81	.61
8	.94	.75
9	.78	.79
10	.81	.59
11	.50	.77
12	.93	.71
13	.95	.67
14	.82	.39
15	.83	.54

those found for the prior strategy.

Discriminant Anzlysis

The Phase III-B data were subjected to a discriminant analysis in which performance on individual items was used to predict class differences. According to Table 20, the presence of classroom differences led to four discriminant functions. In Table 20, the function statistics and the tests for group differences in the residual matrices after extracting each of the discriminant functions are reported (see Tatsuoka, 1971). Table 21 contains the standardized discriminant function coefficients for the first four discriminant functions. Because the directions of the coefficients vary within a function, scales were formed by adding and subtracting items (using items whose standardized discriminant coefficients was larger in absolute magnitude than .40). Scales were formed so that the means on the scales were positive.

Table 22 contains the intraclass correlations for the scales and their regression equations. Again, the intraclass correlations are no higher than the intraclass correlation for the total scale. The discriminant functions did lead to more sensitive indicators of class differences than the total test, because the importance of the within-class variables, most notably the pretest, was

Table 20. BTES fraction items used to discriminate between classes - tests of significance of the first ten functions.

				RESIDUAL STATISTICS		
Function	Eigenvalue	Percent of Variance	Canonical Correlation	Wilks' Lambda	χ^2	df
0				.01	500.57**	300
1	1.46	21.90	.77	.02	407.12**	266
2	1.20	18.08	.74	.04	325.02**	234
3	1.08	16.29	.72	.09	248.71*	204
4	.86	12.96	.68	.17	184.09	176
5	.60	9.06	.61	.27	135.06	150
6	.38	5.74	.53	.38	101.44	126
7	.31	4.59	.48	.49	73.74	104
8	.20	3.85	.41	.59	54.55	84
9	.18	2.65	.39	.70	37.70	66
10	.13	1.91	.34	.78	25.29	50

** significant at α=.0001.

* significant at α=.05.

Table 21. BTES fraction items used to discriminate between classes - standardized canonical discriminant function

	Function			
Item	1	2	3	4
1	-.30	.59	-.16	-.37
2	.79	-.12	.79	.35
3	.54	-.28	.11	.04
4	.43	-.01	-.09	.06
5	.13	.19	.34	.05
6	-.18	-.00	-.45	-.04
7	.09	-.00	-.14	-.53
8	.36	.26	-1.04	.09
9	-.32	.32	.58	-.12
10	-.29	.50	.25	.12
11	-.54	.16	.34	-.19
12	.70	-.10	-.22	.34
13	.22	-.07	.09	.32
14	-.65	-.12	-.14	.64
15	-.23	-.01	.02	.12

Table 22. Scales formed from standardized discriminant functions regressed on the same pretest and the ALT variables.

	Unstandardized				Standardized			
Discriminant Function Number	1	2	3	4	1	2	3	4
Intraclass Correlation	.50	.45	.42	.43				
Between-Class								
Pretest	.53 (1.01)	.37 (.83)	.24 (.33)	.70 (.76)	.28	.23	.09	.21
Allocated Time	-.02 (.34)	-.00 (.10)	-.01 (.30)	-.03 (.58)	-.16	-.04	-.15	-.27
Engagement Rate	-1.36 (.66)	.35 (.26)	1.01 (.86)	-.01 (.00)	-.26	.09	.35	-.00
Low Error Rate	-1.33 (.42)	-1.29 (.62)	.63 (.35)	.48 (.23)	-.17	-.22	.15	.09
High Error Rate	-6.47 (.58)	.40 (.05)	-3.30 (.52)	-5.43 (.74)	-.17	.01	-.16	-.21
Within-Class								
Pretest	.03 (.34)	.18 (1.79)	-.09 (.87)	.07 (.56)	.04	.19	-.09	.06
Allocated Time	.02 (1.12)	.02 (1.73)	.00 (.20)	.01 (.37)	.20	.28	.04	.07
Engagement Rate	.60 (.93)	.47 (1.43)	-.45 (1.24)	-.33 (.76)	.15	.16	-.20	-.12
Low Error Rate	1.08 (1.29)	.06 (.10)	-.20 (.43)	-.79 (1.45)	.19	.01	-.07	-.21
High Error Rate	2.39 (1.13)	-1.29 (.92)	-.20 (.17)	.13 (.09)	.13	-.09	-.02	.01
Constant	1.60	.56	.43	.77				
R^2	.18	.35	.12	.19				

[a] t-statistics within parenthesis - between-class df=15, within-class df=96.

greatly reduced. However, the scales do not seem to be more sensitive to instructional differences. Most of the classroom differences are accounted for by the pretest. In many cases, the between-class coefficients for the ALT variables actually decreased. In addition, subtracting test items in building a scale is not an alternative which researchers would normally practice.

To compare the stability of the discriminant function across samples, the discriminant function for the pilot data was calculated. However, as can be seen from Table 23, only two functions were defined in the pilot data. In addition, the two functions in Table 24 are highly dissimilar from the four discriminant functions in Table 21. Thus, it is concluded that the discriminant functions are not stable across samples.

Between-Class Item-Allocated Time Correlations

As with all the prior techniques, the selection technique used here does not lead to a content cohesive group of items. The correlations, in Table 25, are all positive as might be expected. Classes with more allocated time tend to perform higher on each item. The intraclass correlations for the scales, in Table 26, are again of the same magnitude as the intraclass correlations for the total scale.

Table 23. BTES fraction items used to discriminate between classes in the pilot study.

				Residual Statistics		
Function	Eigenvalue	Percent of Variance	Canonical Correlation	Wilks' Lambda	χ^2	df
0				.29	117.75**	60
1	.67	43.18	.63	.48	69.63*	42
2	.56	36.06	.60	.74	27.93	26
3	.19	12.05	.40	.89	11.30	12
4	.13	8.25	.34			

Table 24. BTES pilot study - standardized discriminant function coefficients.

	Function	
Item	1	2
1	-.14	.93
2	.39	-.08
3	-.12	-.24
4	.22	-.54
5	-.45	.13
6	-.39	.29
7	-.15	-.31
8	.70	-.01
9	-.16	.38
10	.52	-.48
11	-.07	-.67
12	-.01	.12
13	.66	.32
14	-.12	.16
15	.18	.26

Table 25. Between-class correlations of allocated time with fraction items.

	Average Between-Class Correlations	
Item	Pilot	Final
1	.81	.34
2	.98	.34
3	.32	.10
4	.66	.42
5	.80	.59
6	.78	.19
7	.52	.12
8	.86	.40
9	.85	.62
10	.59	.38
11	.33	.60
12	.70	.41
13	.93	.48
14	.87	.59
15	.89	.67

Table 26. Five and ten item scales formed from between class allocated time - item correlation regressed on the same pretest and ALT variables.[a]

	REGRESSION COEFFIICENTS			
	Unstandardized		Standardized	
Cutoff				
Intraclass Correlation	.44	.47	.44	.47
Between Class				
Pretest	.39 (.78	.28 (.51)	.19	.12
Allocated Time	.12 (.80)	.08 (1.04)	.28	.37
Engagement Rate	1.93 (.44)	.54 (.22)	.13	.06
Low Error Rate	-2.56 (.37)	-.66 (.18)	-.11	-.05
High Error Rate	-1.34 (.06)	-.19 (.02)	-.01	-.00
Within Class				
Pretest	.38 (2.91)	.38 (3.06)	.28	.28
Allocated Time	.05 (1.00)	.01 (.58)	.14	.08
Engagement Rate	1.49 (1.09)	1.32 (1.76)	.13	.21
Low Error Rate	2.99 (1.61)	1.60 (1.58)	.18	.18
High Error Rate	-6.31 (1.38)	-4.66 (1.88)	-.12	-.16
Constant	-1.36	-.10		
R^2	.55	.55		

[a] t-statistics within parenthesis - betw-en class df=15, within class df=96.

The regression equations show that this selection technique is effective in building indicators more sensitive to instructional differences. Allocated time assumes a more important role in predicting achievement differences between classes. Furthermore, the within-class effects of allocated time and the pretest are reduced.

Again the cross-sample stability of the index is examined by comparing the five and ten item scales (see Table 25) formed in the two samples. From Tabl 27, it is apparent that the between-class item-allocated time correlation is not stable across samples. Having only two (seven) items in common on two five (ten) item scales is likely to occur randomly (see Table 15).

Summary

Two general conclusions can be made from the analyses reported in this chapter. First, it is apparent that none of the item selection strategies are stable across the samples used in this study. Second, the scales formed within a sample are differentially sensitive to both within-class and between-class variables. While the proportion of between-class variation does not vary widely from scale to scale, different selection techniques lead to different scales which are sensitive to different

Table 27. Overlap of subtests formed from the pilot study and the final study using the between-class correlation of allocated time with the items.

	Top 5		Top 10	
Item	Pilot	Final	Pilot	Final
1			x	
2	x		x	
3				
4				X
5		x	x	x
6			x	
7				
8	x		x	x
9		x	x	x
10				x
11		x		x
12			x	x
13	x		x	x
14	x	x	x	x
15	x	x	x	x
Overlap (number of items)	2		7	

variables.

Each of the five selection techniques produced some similar results. None of the newly formed scales exhibited a higher proportion of between-class variation than the original 15-item selection. Yet, most of the scales behaved differently than the total test in relation to the ALT variables and the pretest. Each of the techniques were effective in reducing the magnitude of the effects of at least some of the variables within classes. In addition, each technique led to a rise in the magnitude of the standardized regression coefficient of at least some of the variables between classes. However, only the average between-class item correlations and the between-class item-total correlations yielded the same results. Some strategies might be more useful for different purposes than others.

The last technique, based on the between-class correlation of the items with allocated time, was effective in increasing the effects of between-class differences in allocated time, in addition to reducing the within-class effects of the pretest and allocated time. The other strategies may lead to an increase in the effect of the variables between classes, but it not known or predictable which effect will be increased. Thus, when the effect of a known variable is desired, that variable could be used in the selection of items. The selection technique would in turn lead to a scale that is more sensitive to the variable of interest.

While the other remaining techniques will increase the sensitivity of the scales to differences between classes, it is not clear whether the variables affected will be a measure of instructional differences (e.g., the ALT variables), and/or a measure of common background or entering ability. Yet, some of the selection techniques have different advantages to the researcher. First, the intraclass correlation or between group item-total correlation are more appealing because they are both simpler, or with the use of a computer, cheaper than the discriminant function.

Overall, the between-class item-total correlation appears to be the most useful of the first four techniques.[2] The discriminant analysis has a number of problems. First, discriminant analysis with discrete variables is not feasible for most researchers at present. Also, even ignoring the possible bias from assuming normally distributed variables, application of the discriminant function has another major problem. The negative signs of the discriminant function coefficients suggest that the best possible discrimination is achieved by adding and subtracting test items. Thus, getting some items right counts positively on a scale, while getting other items right counts negatively on the same scale. Naturally, this leads to problems in interpreting the scale, as well as its relationship to other variables.

Finally, the between-class item-total correlation selection technique is preferrable to using the item intraclass correlations alone because the more stringent criteria led to a more sensitive measure of between-group differences for the item-total correlation, but not for the intraclass correlations. Thus, the magnitude of the between-class regression coefficients increased when the top five items were selected instead of the top ten from the item-total correlations. The coefficients for allocated time, engagement rate, and low error rate increased between classes. In contrast, the between-class coefficients for allocated time, and engagement rate were lower for the five items with the highest intraclass correlations than for the top ten items. In other words, the relationship of the scales to the between-class variables is increased when a more stringent criterion is used for item selection with the between-class item-total correlation strategy, but not with the intraclass correlation strategy.

CHAPTER 5

ANALYSIS OF PATTERNS OF ITEM RESPONSE

In this chapter, patterns of both correct and incorrect responses to test items will be analyzed for classroom differences. First, the correct patterns of response will be examined for the effects of topic coverage and emphasis on achievement item data. Next, the patterns of incorrect responses will be examined for common errors within an instructional group.

Analysis of Patterns of Correct Item Response

A total score gives an indication of overall achievement. However, there are numerous ways of achieving the same score. As Harnisch and Linn (1981) point out, there are 184,756 possible patterns for getting a score of 10 on a 20-item test. Clearly, not all patterns will occur in practice, nor will all of the patterns which occur signify any real differences. Sato (1980) suggested that an index would be useful for diagnosing student response patterns. This index was intended to show the effects of guessing or careless errors. Yet, differences in curriculum coverage and emphasis might also cause different student response patterns between classes. These systematic differences shoulu also be detectable and hopefully, distinguishable from the unsystematic individual differences characterized by guessing and carelessness. In this section the differences in student response patterns between classes is explored using the BTES Phase III-B data, as a means for identifying systematic differences in curriculum coverage and emphasis.

Posttest Differences in Patterns of Item Response

Sato's work (1980) serves as a starting point for this investigation. His S-P Chart is a matrix of student responses by test items. The matrix is transformed so that the items are the columns arranged in descending order of difficulty, (hardest items in the left most column). The rows are the students arranged in descending order of total test score. A revised version of Sato's S-P Chart intended to examine differences between classes is presented in Table 28. The rows are the classes arranged in descending order of mean achievement. The columns are the items in descending order of difficulty. In this table, a classroom which displays the expected pattern of test item performance (i.e., correct on easy items, incorrect on hard items) would have higher numbers in the left-most columns and lower numbers in the right-most columns. Thus, such a class would do better on

Table 28. Number of correct responses to each item by class on occasion C witn items arranged in descending order of item difficulty.

		ITEMS																	
Class	N	1	2	6	4	7	11	10	3	14	12	5	15	13	8	9	Achievement Mean	Caution Index Mean	SD (Caution Index)
27	6	6	6	6	6	6	4	5	5	4	4	6	3	3	5	6	12.50	.27	.30
11	6	6	6	5	5	5	6	5	5	5	6	5	3	5	3	4	12.33	.27	.32
12	6	6	4	6	5	6	5	6	4	6	4	3	3	3	4	3	11.33	.79	.74
24	6	6	6	6	4	6	5	6	3	4	3	3	4	3	4	3	11.00	.39	.27
23	6	6	5	4	6	4	5	6	3	6	5	2	2	2	3	3	10.33	.46	.30
1	5	5	5	5	4	5	3	2	4	2	5	1	1	3	5	1	10.20	.42	.15
8	5	5	5	3	5	4	5	1	4	3	4	3	2	1	0	0	9.00	.30	.20
6	5	5	5	4	5	2	3	2	3	3	3	0	1	2	1	1	8.00	.26	.12
9	5	4	2	4	4	3	3	4	0	4	3	2	2	2	2	1	8.00	.65	.37
5	6	6	6	6	3	5	5	5	3	0	1	2	1	0	0	2	7.50	.27	.15
18	6	4	6	1	5	0	4	1	3	6	5	3	3	4	0	0	7.50	.66	.16
10	6	5	5	4	3	3	3	4	3	1	2	4	1	3	0	1	7.00	.61	.62
26	5	5	5	4	2	2	2	1	1	2	1	1	2	2	1	1	6.40	.40	.42
21	9	7	7	6	5	5	2	1	6	5	3	1	2	2	1	1	6.00	.47	.43
14	6	3	3	4	4	4	2	3	4	1	1	1	0	0	1	1	5.33	.52	.60
19	6	2	2	4	4	5	2	2	6	0	4	1	0	0	0	0	5.33	.52	.43
4	6	4	4	3	3	2	2	3	2	2	1	1	1	1	1	1	5.17	.37	.21
17	6	5	5	4	4	4	1	4	1	1	0	0	0	1	0	0	5.00	.09	.09
25	5	4	4	2	2	2	1	3	2	2	1	2	0	0	0	0	5.00	.34	.43
3	6	2	2	2	1	2	3	2	2	3	2	2	2	2	1	1	4.83	.38	.49
16	5	1	1	2	1	1	2	0	1	5	0	0	3	0	0	0	2.83	.74	.60
Total		.79	.76	.69	.66	.62	.55	.54	.53	.53	.47	.35	.33	.32	.26	.24			

easier items (highest means) and worst on harder items (lowest means). While no class in Table 28 has exactly the same ranking of item difficulty as the total sample, some classes do exhibit a pattern of performance similar to the total sample. For example, classrooms 3 and 11 do not have any large differences between two adjacent columns.

However, some classes clearly exhibit patterns of item responses different from the total sample. Classroom 19 has a six point difference between two items with the same overall difficulty (3 and 14), and some surprisingly low scores on some of the easiest items (only 2 correct on items 1 and 2). Classroom 16 does well on items 14 (5 correct) and 15 (3 correct) despite an overall low achiavement level. Classroom 18 also has what appears to be a random pattern of responses. Students from this class do poorly on three of the seven easiest items (6,7, and 10), but perform well on the next six items.

While some of the cited patterns of item response in Table 28 may be random (e.g., differences due to guessing or carelessness), others may be attributed to differences in classroom instruction. Within the area of fractions, many different skills may be taught (e.g., addition, subtraction, multiplication, etc.). Furthermore, these differences are often reflected in the items of a test and not in the total test. Thus, differences in topic coverage and emphasis may be reflected in the achievement items.

In the fifteen item test under examination, four different fractions operations can be identified. The first five items cover the subtraction of fractions. The second five items cover the addition of fractions. Items 11, 12, and 13 require the student to recognize equivalent forms of fractions, and items 14 and 15 involve solving fractions equations for an unknown numerator or denominator.

To facilitate the discussion of possible differences in topic coverage and emphasis, the columns of Table 28 are rearranged into the original order of items. Thus, the lines dividing the columns in Table 29 into four groups represent the different content areas covered by the fractions test. In Table 28, some classrooms stand out for their apparent differences from the total sample in their topic coverage or emphasis. For example, the lowest achieving class (16) does not do well in subtraction (items 1-5), addition, (6-10), nor equivalences (11-13), but does seem to have effectively learned algebraic manipulation (14 and 15).

When class 16 is compared with classroom 3, two very

Table 29. Number of correct responses within class on BTES fraction items on occasion C with items grouped by content area.

ITEM

Class	N	Subtraction					Addition					Equating			Algebraic Manipulation		Achievement Mean	Caution Index Mean	SD (Caution Index)
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15			
27	6	6	6	5	6	6	6	6	5	6	5	4	4	3	4	3	12.50	.27	.30
11	6	6	6	5	5	5	5	5	3	4	5	6	6	5	5	3	12.33	.27	.32
12	6	6	4	4	5	3	6	6	4	3	6	5	4	3	6	3	11.33	.79	.74
24	6	6	6	3	4	3	6	6	4	3	6	5	3	3	4	4	11.00	.39	.27
23	6	6	5	3	6	2	4	4	3	3	6	5	5	2	6	2	10.33	.46	.30
1	5	5	5	4	4	1	5	5	5	1	2	3	5	3	2	1	10.20	.42	.15
8	5	5	5	4	5	3	3	4	0	0	1	5	4	1	3	2	9.00	.30	.20
6	5	5	5	3	5	0	4	2	1	1	2	3	3	2	3	1	8.00	.26	.12
9	5	4	2	0	4	2	4	3	2	1	4	3	3	2	4	2	8.00	.65	.37
5	6	6	6	3	3	2	6	5	0	2	5	5	1	0	0	1	7.50	.27	.15
18	6	4	6	3	5	3	1	0	0	0	1	4	5	4	6	3	7.50	.66	.16
10	6	5	5	3	3	4	4	3	0	1	4	3	2	3	1	1	7.00	.61	.62
26	5	5	5	1	2	1	4	2	1	1	1	2	1	2	2	2	6.40	.40	.42
21	9	7	7	6	5	1	6	5	1	1	1	2	3	2	5	2	6.00	.47	.43
14	6	3	3	4	4	1	4	4	1	1	3	2	1	0	1	0	5.33	.52	.60
19	6	2	2	6	4	1	4	5	0	0	2	2	4	0	0	0	5.33	.52	.43
4	6	4	4	2	3	1	3	2	1	1	3	2	1	1	2	1	5.17	.37	.21
17	6	5	5	1	4	0	4	4	0	0	4	1	0	1	1	0	5.00	.09	.09
25	5	4	4	2	2	2	2	2	0	0	3	1	1	0	2	0	5.00	.34	.43
3	6	2	2	2	1	2	2	2	1	1	2	3	2	2	3	2	4.83	.38	.49
16	6	1	1	1	1	0	2	1	0	0	0	2	0	0	5	3	2.83	.74	.60
Total		.79	.76	.53	.66	.35	.69	.62	.26	.24	.54	.55	.47	.32	.53	.33			

different interpretations emerge. Some students in classroom 3 may have mastered the material, but the rest of the class does not understand the material. (Three students had total scores of zero.) This phenomenon could be the result of ineffective instruction (high achievers might learn the material in spite of the instruction), or the instruction may be aimed only at some of the students (e.g., differential topic coverage within the class). In either case, there are marked instructional differences from classroom 16, where there appears to be effective instruction, but only in one area.

Another example of possible instructional differences is suggested by a comparison of classes 5 and 18. In this case, an examination of total test scores would show no difference between the two classes because mean achievement is exactly equal in the two classes. (Both classes have means at the midpoint of the scale (7.5).) However, an examination of the item response patterns show different phenomena occurring in the two classes. Classroom 18 does well in three of the content areas, but the students cannot add fractions (items 6-10). In contrast, classroom 5 can add fractions, but cannot perform the algebraic manipulations necessary for problems 14 and 15 nor the equivalences in problems 12 and 13.

Caution Index. While Table 29 is useful for discussing classroom differences in coverage and emphasis in this study, large-scale evaluations typically involve too many items and/or too many groups (classes, schools, etc.) to make generating such a table practical. Thus, an index of response patterns is needed to identify groups with peculiar test behavior. Sato's caution index (see Chapter 2) has been used at the student level to measure the anomalousness of response patterns. Sato proposed that the individual response pattern should be examined when the index was greater than or equal to .5. The cutoff of .5 might also be used on the class mean of the caution index. Thus, classes with a mean caution index above .5 would be examined for the atypical response pattern.

Sato's explanations for the high index might still be applicable. That is, many of the members of a high achieving classroom might be making careless mistakes. Or the members of a low achieving classroom may be getting correct answers from guessing. However, an alternative explanation might be that curriculum coverage and emphasis are causing the classroom to answer items atypically from the rest of the sample. Thus, classrooms 16 and 18 should be high on the caution index.

The student response patterns and their caution indices are reported in Appendix D. The class mean and standard deviations on the caution index are reported in Tables 28 and 29. The mean caution indices yield seven classes with a mean greater than .5. Classrooms 12, 9, 18, 10, 14, 19, and 16 had mean caution indices of .79, .65, .66, .61, .52, and .74, respectively. The high class means on the caution index combined with the individual patterns of item response in Appendix D suggest three possible explanations for an atypical response pattern. The first two reasons for an atypical response pattern are guessing and carelessness, as Sato suggested. Apparently, the class mean index is affected by these two factors when they are based on only six cases. Thus, classes 12 and 9 have a high mean caution index because of carelessness. Carelessness is assumed when there are no consistent patterns of responses within the class and achievement is high. In addition, classes 10 and 19 have a high mean caution index because of guessing. Guessing is assumed when there is no consistent pattern of responses and achievement is low. Finally, a third possible reason for the high mean caution index is differences in topic coverage and emphasis. As discussed above, classrooms 16 and 18 were high on the caution index, as expected, because of apparent instructional differences from the total sample.

In addition to the three possible reasons for atypical responses, these reasons may occur simultaneously in the same class. For example, in class 14 the higher achievers made careless errors, while the low achievers correctly answered some problems by guessing. Also, classroom 16 appears to be affected by guessing on the first thirteen items of the test.

Because the high caution index may be due to random processes (guessing or carelessness) or to instructional differences, the index is limited in its usefulness for our purposes. Yet a difference between the unsystematic causes and instructional differences may be indicated by the standard deviation of the caution index. When the atypical response is the same throughout the class (as might be expected in the case of instructional differences), the patterns of response would be fairly similar and the standard deviation of the caution index would be small. However, when the patterns of response within a class are highly dissimilar (as might be expected from guessing or carelessness), the standard deviation of the caution index would be high. Thus, the ratio of the index over the standard deviation of the index may be more useful than just the mean. This also raises questions about the psychometric properties of the caution index. The standard deviations in Tables 28 and 29 are low as expected for 18, but not for class 16. However, class 16, as mentioned earlier, is affected by guessing. Thus, the only class (18) with a high mean caution index that is apparently affected by instruction and not guessing or carelessness is also the only class with a low standard deviation on the caution index in conjunction with a high mean caution index.

Pretest Differences in Patterns of Item Response

While it is important to examine differences in patterns of response after instruction, different patterns of response prior to instruction lead to different conclusions about the effects of instruction. That is, conclusions about the effects of instruction are stronger when the change from achievement prior to instruction can be analyzed. Consequently, Table 30 provides the analog of Table 29 prior to instruction. That is, data from the same 123 students in Table 29 are reported in Table 30. Again, each entry in the table represented the number of students in a given class ooms (row) that answered a given item (column) correctly. In addition, the class means in achievement, and the class means and standard deviations on the caution index are reported in Table 30.

From Table 30, eight classes (27, 12, 24, 8, 18, 10, 21 and 3) have a mean caution index greater than or equal to .5. However, unlike the patterns of response after

Table 30. Number of correct responses within class on BTES fraction items on occasion B with items grouped by content area.

		ITEM																	
		Subtraction					Addition					Equating			Algebraic Manipulation		Achievement Mean	Caution Index Mean	SD (Caution Index)
Class	N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15			
27	6	2	3	5	3	3	3	3	1	0	3	4	2	0	2	2	6.00	.59	.35
11	6	6	6	6	4	2	5	5	2	1	4	5	5	3	6	4	10.67	.30	.41
12	6	3	2	3	2	1	3	2	0	0	3	3	2	1	2	3	5.00	.67	.59
24	6	3	3	3	3	0	5	5	2	2	3	4	4	4	3	2	7.67	.66	.56
23	6	3	3	4	5	0	1	1	1	1	3	3	1	1	1	2	5.00	.40	.34
1	5	5	3	3	2	0	4	2	1	0	1	1	1	1	1	1	5.20	.36	.45
8	5	1	3	1	1	0	2	0	1	1	3	2	1	0	1	0	3.40	.61	.51
6	5	4	4	3	4	0	4	4	2	2	4	2	1	1	3	3	8.20	.32	.35
9	5	5	5	1	4	0	5	5	1	1	2	2	3	1	2	2	7.80	.41	.20
5	6	2	0	3	1	1	1	0	0	0	2	4	0	0	0	0	2.33	.47	.35
18	6	2	2	1	2	0	1	1	0	0	0	1	0	1	2	2	2.50	.53	.58
10	6	3	2	4	2	3	4	3	0	1	2	1	1	1	3	0	5.00	.65	.23
26	5	1	1	1	1	0	1	0	0	0	0	1	0	0	1	0	1.40	.20	.44
21	9	3	4	3	5	2	1	1	0	0	1	1	0	2	2	1	2.89	.67	.61
14	6	2	1	3	0	1	1	0	0	0	0	0	0	1	1	0	1.67	.30	.38
19	6	6	6	2	4	0	3	4	1	0	3	1	1	0	1	1	5.50	.24	.16
4	6	2	1	3	2	0	2	1	0	0	1	2	1	1	1	1	3.00	.44	.40
17	6	4	2	3	3	0	2	2	1	0	2	0	1	0	1	0	3.50	.17	.14
25	5	3	3	2	1	0	3	2	0	0	2	1	2	1	2	1	4.60	.46	.18
3	6	0	0	2	0	0	2	2	0	0	0	2	1	0	0	0	1.50	.50	.53
16	6	0	1	2	1	0	1	0	1	0	1	0	0	0	1	0	1.33	.21	.41
Item Means		.49	.45	.47	.40	.11	.43	.34	.11	.07	.32	.32	.22	.15	.29	.20			
Order		1	3	2	5	13	4	6	13	15	7	7	10	12	9	11			

instruction, none of the eight classes stand out as being different from the total sample. Instead, the high caution index appears to be the result of guessing and/or carelessness. In addition, the standard deviation of the caution index within classrooms is higher prior to instruction. The standard deviation of the index prior to instruction is an average of .04 higher than after instruction. Also, the standard deviations are higher in two thirds of the classes prior to instruction. Thus, instruction has the effect of making the response patterns within a group more uniform (reduced variability).

Pretest-Posttest Change in Patterns of Item Response

Finally, the change in class achievement is reported in Table 31. Each entry in Table 31 represents the differences between the entry in Table 29 and the same entry in Table 30. Also, the mean allocated time in minutes per week is reported in Table 31. The mean achievement has gone up in 19 of the 21 classes indicating an instructional effect.

Some classes have more or less of an instructional effect. For example, classroom 19 has a drop in achievement, especially on items 1 and 2, which is reasonable, given that no time was devoted to fractions. Also, the earlier conclusions about classroom 16 seem

Table 31. Change in achievement by class from occasion B to occasion C with items grouped by content area.

		ITEM																
		Subtraction					Addition					Equating			Algebraic Manipulation			
Class	N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Change in Mean Achievement	Allocated Time
27	6	4	3	0	3	3	3	3	4	6	2	0	2	3	2	1	6.50	114.10
11	6	0	0	-1	1	3	0	0	1	3	1	1	1	2	-1	-1	1.66	71.44
12	6	3	2	1	3	2	3	4	4	3	3	2	2	2	4	0	6.33	86.89
24	6	3	3	0	1	3	1	1	2	1	3	1	-1	-1	1	2	3.33	45.74
23	6	3	2	-1	1	2	3	3	2	2	3	2	4	1	5	0	5.33	88.70
1	5	0	2	1	2	1	1	3	4	1	1	2	4	2	1	0	5.00	26.06
8	5	4	2	3	4	3	1	4	-1	-1	-2	3	3	1	2	2	5.60	25.67
6	5	1	1	0	1	0	0	-2	-1	-1	-2	1	2	1	0	-2	-.20	7.19
9	5	-1	-3	-1	0	2	-1	-2	1	0	2	1	0	1	2	0	.20	49.19
5	6	4	6	0	2	1	5	5	0	2	3	1	1	0	0	1	5.17	77.25
18	6	2	4	2	3	3	0	-1	0	0	1	3	5	3	4	1	5.00	118.58
10	6	2	3	-1	1	1	0	0	0	0	2	2	1	2	-2	1	2.00	32.84
26	5	4	4	0	1	1	3	2	1	1	1	1	1	2	1	2	5.00	45.53
21	9	4	3	3	0	-1	5	4	1	1	0	1	3	0	3	1	3.11	16.98
14	6	1	2	1	4	0	3	4	1	1	3	2	1	-1	0	0	3.67	29.14
19	6	-4	-4	4	0	1	1	1	-1	0	-1	1	3	0	-1	-1	-.17	0.0
4	6	2	3	-1	1	1	1	1	1	1	2	0	0	0	1	0	2.17	22.37
17	6	1	3	-2	1	0	2	2	-1	0	2	1	-1	1	0	0	1.50	4.21
25	5	1	1	0	1	2	-1	0	0	0	1	0	-1	-1	0	-1	.40	2.71
3	6	2	2	0	1	2	0	0	1	1	2	1	1	2	3	2	3.33	3.34
16	5	1	0	-1	0	0	1	1	-1	0	-1	2	0	0	4	3	1.50	51.52

valid. That is, there is instruction (51.52 minutes per week) which is effectively teaching algebraic manipulations (items 14 and 15). Finally, classrooms 5 and 18 are learning subtraction. But class 5 is learning addition, while class 18 studied algebraic manipulations and recognizing equivalent forms of a fraction.

Apparently, topic coverage and emphasis plays a key role in classroom achievement. Classroom 18, while having a higher allocated time to fractions (118.58) than any other class, was only in the middle of the distribution in achievement. However, if class 18 had covered fractions addition and had scored the same as on the other three topics in the test, then their mean achievement would be 10.75 or 3.25 points higher, which would have placed them among the top five classes overall.

Caution Index and Possible Predictors

Given that real differences do exist in instruction that can affect an index of atypical response patterns, the question remains whether this index relates to achievement and instructional variables. In Table 32, the caution index for individual students is regressed on the ALT variables and a pretest, using the contextual effects model (i.e., entering student-level variables and group means simultaneously). The F statistic for the whole

Table 32. Caution index on occasion C predicted from ALT variables and pretest (i.e., fractions test occasion B).[a]

	REGRESSION COEFFICIENTS	
	Unstandardized	Standardized
Between Class		
Pretest	-.03 (.50)	-.18
Allocated Time	.00 (.15)	.08
Engagement Rate	.18 (.20)	.09
Low Error Rate	-.38 (.28)	-.12
High Error Rate	-.43 (.09)	-.03
Within Class		
Pretest	.01 (.75)	.11
Allocated Time	.00 (.19)	.04
Engagement Rate	-.11 (.39)	-.07
Low Error Rate	.24 (.66)	.11
High Error Rate	.34 (.37)	.05
Constant	.46	
R^2	.04	

[a] t-statistics within parenthesis - between class df=15, within class dt=96.

equation and the R^2 indicate that the index is unrelated to the ALT variables and the pretest ($F=.45$, $df=10, 111$, and $R^2=.04$). In addition, the analysis of variance in Table 33 shows that the class to class differences on the caution index are not significant ($F=.31$, $df=20, 102$).

While there are no significant differences between classes on the caution index, the index still may be a useful tool for measuring differences in curriculum coverage and emphasis. However, the effects of guessing and careless mistakes make the index more difficult to interpret. This is especially true since greater variability in guessing and carelessness should be expected within classes than between classes.

Still, a number of findings indicate the usefulness of the index. First, the high index for classes 16 and 18 in Table 29 is substantively meaningful. Second, floor or ceiling effects on a test would reduce the possible response patterns. For example, class 27 in Table 29 may vary in its patterns of response on a more difficult test; however, on this test no information about what was not covered can be gained. Similarly, when a test is too difficult, n information is gained about what is covered and emphasized. When a student gets all of the items correct or all of the items incorrect, no information is

Table 33. ANOVA - Caution Index by class.

Source	SS	df	MS	F
Class	3.70	20	.19	1.15
Residual	16.41	102	.16	
Total	20.11	122		

available from the response pattern. Thus, information might be lost about classroom differences. In fact, the intraclass correlation of the caution index after instruction excluding achievement scores of 0 or 15 (η^2=.25) is greater than the intraclass correlation of the caution index for the total sample (η^2=.18).

Finally, after excluding the cases with no information because of an achievement score of 15 or 0, the index was compared on all three occasions -- prior to instruction, after instruction, and after the summer break. As expected, the intraclass correlation was lower after the summer break (.14), and higher after instruction (.25) than prior to instruction (.20).

Patterns of Error Response

While patterns of correct responses yield useful information about classroom differences, the incorrect responses may also be used to diagnose classroom differences. An error which occurs only once or twice on a test can be assumed to be merely random. This error could be due to guessing, carelessness, or any other reason. In short, an occasional error gives no information about the reasons for the incorrect response. On the other hand, an error which repeatedly occurs on similar problems can be assumed to be systematic. If a student continually responds in the same incorrect way over similar problems, then it can safely be assumed that the student is following some algorithm to arrive at the answers, but the algorithm or procedure is incorrect. Thus, the low achiever may be from one of two groups. The low achiever may be trying to the best of his (her) ability, but using an incorrect algorithm. Or the low achiever may have no understanding of the material, or is not trying, either of which will lead to random error responses.

Employing an incorrect algorithm, like following a correct algorithm, can be related to individual experiences or to experiences common to the group. That is, the incorrect algorithm may be a result of something the student brings to the class, or the classroom may affect the algorithm used. Individuals responding differently to instruction may come up with different algorithms from their interpretation of the instruction, or an incorrect algorithm may exist prior to instruction which is not altered. On the other hand, students may enter the classroom without any algorithm for a problem (neither correct nor incorrect) and leave with a common procedure for answering the problem which is correct or incorrect. An incorrect procedure may be due to faulty instruction or incorrect transfer of skills from one area to another which is not taught. In either case, valuable information is gained by diagnosing the common incorrect algorithm. The information can then be used to change or add to the instruction to eliminate the incorrect algorithm.

Between-Class Variation in Incorrect Responses

In this section, methodology for diagnosing incorrect algorithms is adapted for detecting classroom-level patterns of incorrect responses. The intraclass correlations for the distractors on each item on the three occasions are reported in Tables 34, 35, and 36. The intraclass correlation for a given distractor is calculated in the same way as the intraclass correlation for a correct response, except the item is coded one when the given distractor is chosen and zero otherwise. The intraclass correlations were calculated only for those distractors which were chosen by 10 or more students overall. It was assumed that the distractors with only a few respondents were chosen randomly. A summary of the distractor intraclass correlations in Tables 34, 35,, and 36 is contained in Table 37.

For the distractors, the same pattern holds as for the correct responses. Post instruction has the highest proportion of between-class variation (.23). However, the

Table 34. Distractor intraclass correlations - occasion B.[a]

Item	1	2	3	4	NR
1	.24	.18	*		.44
2	.19	*		.25	.43
3	*	.15		.30	.35
4	.19			*	.43
5	*	.20	.19	.15	.36
6	*	.31			.48
7	.34	*			.39
8	.28	.16		*	.36
9		.25	*		.37
10		*	.12	.16	.38
11	.19	*	.27		.28
12	.25	.26	*	.23	.31
13	*	.19	.23	.13	.28
14	.17	.11	*		.37
15	.19	.19	.13	*	.34

[a]* indicates the correct repsonse.

Table 35. Distractor intraclass correlations - occasion C.[a]

Item	1	2	3	4	NR
1	.34		*		.43
2		*		.31	.37
3	*	.15		.16	.44
4	.18			*	.44
5	*	.23		.17	.48
6	*	.27			.38
7	.30	*			.38
8	.32	.26		*	.38
9		.31	*	.13	.52
10	.14	*		.35	.35
11	.27	*	.13		.33
12	.12	.22	*	.21	.40
13	*	.32	.13	.19	.39
14	.24	.43	*	.17	.27
15	.19	.21	.20	*	.27

[a]* indicates the correct response.

Table 36. Distractor intraclass correlations - occasion D.[a]

Item	1	2	3	4	NR
1	.29		*		.21
2		*		.27	.21
3	*	.13		.20	.29
4	.34			*	.36
5	*	.15		.25	.29
6	*	.26			.21
7	.22	*			.35
8	.30	.18		*	.24
9		.19	*		.35
10		*		.24	.36
11	.14	*			.22
12		.21	*		.29
13	*	.19	.15	.19	.19
14	.17	.13	*		.29
15	.27	.22		*	.26

[a] * indicates the correct response.

Table 37. Mean intraclass correlations for distractors and no response.

	Occasion		
	B	C	D
Distractors	.21	.23	.21
No Response	.37	.39	.27
Combined	.26	.28	.24

proportion of between-class variation after the summer break (.21) is no lower than prior to instruction (.21).

The intraclass correlations for the no responses is also highest after instruction. However, the percent of between-class variation is lowest after the summer break. The high percent of between-class variation in no response is probably because of a teacher's instructions to either guess or leave blank an unknown answer. Thus, after the summer break, the students from a given class may be spread into new and different classes. The new teachers may not have affected achievement nor incorrect responses yet. However, instructions on whether to guess or leave blank answers to unknown problems will immediately affect the students. Hence, the between-class variation (based on last year's classes) is immediately reduced for the no response alternative.

Diagnosing Errors Prevalent in Individual Classes

From the preceding discussion, it appears that individual distractors are affected by class membership. But the question still remains of whether the errors existing in individual classrooms can be diagnosed, and if so, which classrooms should be examined. Two types of classes are likely candidates for examination of their error response patterns. Low achieving classes obviously may be using incorrect algorithms. In addition, classrooms high on the caution index may have an incorrect algorithm for some subset of problems which led to an atypical response pattern, but not low achievement overall.

The distractors prevalent in a given class are reported in Table 38. An examination of the individual classes and their distractors gives information about what instructional changes might affect achievement. For example, the two lowest achieving classes (16 and 3) seemed to have emphasized the no response alternative.[3] Since most classrooms were credited for correct guessing, the mean achievement for classes 3 and 16 might have been raised by guessing. In fact, guessing, assuming purely random responses (i.e., probability of correct answer equals .25), would have raised the average achievement for classrooms 3 and 16 by 1.98 and 2.13, respectively. This almost doubles the achievement score for classroom 16.

In addition to the no response alternative, a number of logical errors in classrooms can be found. However, this is not true for the high achieving classrooms with careless errors (e.g., classes 9, and 12). Thus, the mean caution index is not useful when it reflects carelessness. The most important classes for this sort of analysis seem

Table 38. Common distractors chosen by half or more of the students within each class on occasion C.

	ITEM																													
Class	1		2		3		4		5		6		7		8		9		10		11		12		13		14		15	
	D*	N*	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N
27																													1	3
11															1	3													1	3
12																														
24									2	3							2	3												
23			2	3					2	4															2	3				
1																	2	3												
8															2	3	2	5	4	4									1	3
6									2	4			1	3	1	3	2	4												
9																	2	4												
5							1	3	2	4					2	4							2	3	2	5	1,2	3	3	3
18											2	5	1	6	1	6	2	6	4	5										
10															1	4	2	4					2	3			1	3	2	3
26															1	3	2	4			1	3								
21									2	7					1	7	2	8	4	5	1	5								
14	1	3	4	3											1	4	2	4			1	4	4	3	2	6	1	3	2	3
19	1	4	4	4													1	4	2	5					4	3	2	5	1	4
4													1	4	1	4	2	4					4	3						
17									2	3					2	3	2	4			1	3	NR**	3	2	3	NR	3	NR	3
25											2	3			1	5	2	4			1	3			2	3				
3	NR	3	NR	3	NR	3	NR	3	NR	3	NR	3	NR	3	NR	3	NR	4	NR	3	NR	3	NR	3	NR	3	NR	3	NR	3
16	NR	3	NR	3	NR	5	NR	5	NR	5			NR	3	NR	3	NR	3	4,NR	3	NR	4	NR	4	NR	5				

*D=distractor; N=number of cases per class choosing distractor

**NR=no response

to be the low achieving classroom and the classroom with an atypical curriculum sequence (e.g., class 18).

In the case of the individual student, distractors become important when the same incorrect procedure leads to similar responses on similar problems. Likewise, in examining classes, the key is to find the incorrect algorithm which will cause many students from the same class to produce similar incorrect responses for similar problems. This phenomenon can be seen throughout Table 38. For example, the most common error in fractions addition is to add in both the numerator and the denominator. This incorrect algorithm would lead to responses 2, 1, 1, 2, and 4, for items 6, 7, 8, 9, and 10, respectively. An examination of Table 38 reveals that classes 4, 18, 21, and 25 do have multiple respondents on each of these incorrect responses and not on the other incorrect responses. Multiple problems are needed because using only one or two problems would not always lead to the correct conclusion. For example, in classroom 17 students used the same incorrect response on item 9. However, the same incorrect responses on items 6, 7, 8, and 10, had only one respondent using the same algorithm. Thus, the 4 responses on choice 2 of item 9 can only be considered random or at least not relevant to this type of analyses.

Besides adding in both the numerator and denominator always, some classes have that algorithm only when the correct answer appears in some equivalent form. So classrooms 10, 14, 19, and 26 use the same algorithm for items 8 and 9, when the correct answer appears in another form (e.g., 1/3 + 2/3 = 1 and not 1/3 + 2/3 = 3/3). However, the same classrooms use a correct algorithm when the answer is in the same form as the problem (i.e., items 6, 7, and 10).

Besides addition errors, incorrect algorithms are used by multiple members of classes in subtraction and equating. For example, the answer used by classrooms 14 and 19 when no whole numbers are used, is only the numerator (ignoring or dropping the denominator). That is, distractors 1 and 4 are used for items 1 and 2, respectively. Also in equating, classrooms 14, 17, 21, and 25 answer makes sure the difference between the numerator and denominator is the same for the two fractions (e.g., 5/7 = 2/4 = 10/12). That is, distractors 1, 4, and 2 are used on items 11, 12, and 13, respectively. Finally, distractors may be paired in ways that are not immediately obvious to the researcher, but may be systematic. For example, on items 14 and 15, distractors 1 and 2 are paired in classroom 10. In addition, distractors 2 and 1, for items 14 and 15, are paired in classroom 19.

Pretest Differences in Patterns of Error Response

While analyzing error responses within classes can supply valuable information about what errors students are making and thus what adjustments to instruction need to be made, the cause of the common error is still unknown. By using a pretest, the common error can either be diagnosed as a result of the common experiences during the instruction period (i.e., not present during the pretest), or present prior to instruction but not corrected. In Table 39, the distractors chosen by at least half of the members in a class prior to instruction (occasion B) are reported.

From Table 38 and 39, a number of different phenomena can be seen. First, the error found in class 18 on the posttest (adding both the numerator and the denominator) is prevalent in many classes (27, 23, 18, 10, 21, 14, and 4) prior to instruction. However, the error was corrected in the classes to varying degrees. In some classes (27. and 23) the error is virtually eliminated; in others (5 and 26) the error is only present when the problems also require reducing the fractions (items 8 and 9); and in class 18, the error continues unabated after instruction.

Classroom 16 also appears to have effectively learned

Table 39. Common distractors chosen by half or more of the students within each class on occasion B.

	ITEM																													
Class	1		2		3		4		5		6		7		8		9		10		11		12		13		14		15	
	D*	N*	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N	D	N
27	1	3	4	3			1	3			2	3	1	3	1	4	2	6					2	3	2	4				
11									2	3					1	3	2	5												
									3	3					2	3	2	3			1	3	2	3						
24	NR**	3	NR	3	NR	3	NR	3	NR	3					1	3	2	3												
23	1	3	4	3							2	4	1	5	1	5	2	5	4	3									NR	3
1							NR	3	NR	3							2	4	NR	3	1	3								
8					2	3			2	4			1	3	1	3									2	4				
6															1	3	2	3					4	3						
9					4	3			2	4							2	3											2	3
5			4	3					NR	3	2	4	1	5	1	5	2	5					2	4	2	3	NR	3	3,NR	3
18							1	3	2	3	2	4	1	4	1	5	2	5	4	4	NR	3	NR	3	NR	3				
10													1	3	1	3	2	5			3	4	2	4	2	3			1,3	3
26	NR	4	NR	4	NR	4	NR	4	NR	4	NR	4	NR	4	NR	4	NR	4	NR	4	NR	3	NR	3	NR	3	NR	3	NR	3
21											2	7	1	7	1	7	2	8			1	5								
14	1	3	4	3			1,NR	3	NR	3	2	4	1	4	1	4	2	4	4	3	1,NR	3	NR	3	NR	3	NR	3	NR	3
19					4	3			2	4	2	3			1	5	2	5			1	4	1	4	2	3				
4	1	3	4	4							2	4	1	4	1	4	2	4												
17									2	3							2	3			1	4	NR	3	2,NR	3	NR	3	NR	3
25													1	3	1	4	2	5			1	3								
3	1,NR	3	NR	3	NR	3	NR	4	NR	4	NR	3	NR	3	NR	3	NR	3	NR	5			NR	4	NR	4	NR	5	NR	4
16	NR	4	NR	4	NR	4	NR	5	NR	5	NR	5	NR	5	NR	4	NR	5	NR	5	NR	5	NR	5	NR	5	NR	5	NR	5

*D=Distractor; N=Number of cas s per class choosing distractor

** NR=No response

algebraic manipulations. As further evidence for the students' learning, the no response alternative is used on all fifteen items on the pretest, as it was on twelve of the first thirteen items on the posttest. Hence, the high achievement on the algebraic manipulations seems not to be a chance occurrence when the no response alternative is used so frequently instead of guessing. Similarly, the no response alternative is used in classroom 18 for the equivalent fractions. And again there is high achievement after instruction.

Summary

Patterns of response, whether correct or incorrect, provide information about differences in classroom achievement which is not present in the total score. In addition, the use of items seems more useful than the creation of subsets of scores based on common content. This is primarily because the use of items lets the data determine the subsets of items with common behavior rather than the researcher. For example, the fractions subtest used in this chapter appears to have four distinctive groups of fractions items -- addition, subtraction, recognizing equivalent fractions, and algebraic manipulations. However, an examination of the common errors in Tables 38 and 39 seem to indicate that items 1 and 2 are clearly similar, but that error responses in items 1 and 2 are unrelated to items 3, 4, and 5. In fact, using the incorrect algorithm most prevalent for items 1 and 2 (subtracting the numerator and denominator) leads to a correct response for item 3.

Two sources of information were used to examine classroom differences -- incorrect responses (distractors) and correct responses. Each seems useful for a different purpose. The patterns of correct item response are useful to the researcher for analyzing differences in topic coverage and emphasis. However, the analysis of errors seems more useful as an instructional adjunct for making adjustments to the instruction. While the analysis of correct item response may be useful as an instructional adjunct, it is assumed that teachers are aware of their own curriculum sequence. But they may still be unaware of the errors which need to be overcome.

CHAPTER 6

SUMMARY AND CONCLUSIONS

.Overview of Chapters

In this dissertation, several possible empirical strategies for measuring achievement differences between instructional groups were examined. In Chapter 1, the context of the research problem was.outlined and a general statement of the focus of this study was put forth. This study focused on standardized achievement tests and their perceived inadequacy in detecting program effects (lack of instructional sensitivity or program relevance). This concern led to the consideration of certain empirical strategies for insuring greater instructional sensitivity and program relevance.

In Chapter 2, the literature relevant to this study was described. The chapter began with a brief review of the background which led to the question of why instructional sensitivity is an issue in the school effects literature. In addition, the pertinence of the literature on the analysis of multilevel data was discussed. Then literature relevant to our two lines of inquiry were discussed. First, past efforts at constructing more sensitive indicators of group differences, through the use of indices of item discrimination between groups, were reported. Second, procedures for the analyses of patterns of correct and incorrect responses which have been traditionally used at the individual level were discussed as possible measures of group phenomena.

In Chapter 3, the proposed analytical strategies were described. These strategies were divided into two groups. The first group of strategies suggested that five indices of item discrimination between groups might be used to build scales which are more sensitive to between-group differences. The five indices considered for use in test construction were the item intraclass correlations, the item intraclass correlations in conjunction with the between-group item intercorrelations, the correlations of the group means on the scale with the group means on the items, a discriminant analysis, where items are used to discriminate between groups, and the between-group correlation of the items with an instructional variable. Each of the five indices were viewed as possibly useful for selecting items into a scale which would be more sensitive to between-group differences, especially those associated with instructional and program variables.

The second group of analytical strategies proposed in Chapter 3 adapted procedures previously employed at the individual level to measure differences between groups in patterns of item response. First, group patterns of correct item responses were suggested for use in measuring group differences in instructional coverage and emphasis. A group-level version of the index (Sato, 1980) was explored as a possible statistical measure of differential instructional coverage and emphasis. Second, the use of patterns of incorrect item responses as information about group differences was discussed. The difference between random and systematic errors in these patterns was discussed and guidelines for attributing an error group-level phenomena were proposed.

Finally, in Chapter 3, the BTES data set used for the empirical investigation was described. Furthermore, the means of applying the empirical strategies to the data set were further elaborated.

In Chapters 4 and 5, the analytical strategies proposed in Chapter 3 were applied to the fifth grade fractions test from the BTES data. Chapter 4 used the five indices of item discrimination between groups to form scales, and Chapter 5 analyzed the patterns of correct and incorrect item response.

Summary of Results

Chapters 4 and 5 recount a substantial number of specific results from the empirical investigations. Despite limitations in the empirical data considered (to be discussed below), certain trends in results suggest a number of specific interpretations from the data.

Subsets of Group-Sensitive Items

Our investigation of the five alternative indices for selecting items for constructing scales more sensitive to group differences pointed out a number of similarities among the indices. First, with exception of the between-group item-total correlation and the intraclass correlation - between-group correlation, indices selected slightly different sets of items. Second, the scales constructed by all five indices exhibited approximately the same proportion of variation between classes (ranging from .42 to .50) as the total scale (.47). However, the relationships of the scales to the ALT variables (the available measures of instruction) and the pretest varied depending on the index used for item selection. The index based on the between-class correlation of the items with allocated time was most effective in building a scale sensitive to the variable of interest. (In this case, the between-class relationship to allocated time increased.) In contrast, the other techniques may have behaved differently than the total scale, but the differences were not predictable. So if the relationship of some variable (e.g., instruction) to achievement is desired, selecting items on the basis of their relation to that variable appears to be the most effective strategy among those considered.

When there is no single variable of interest, one of the remaining techniques might be considered. The technique least likely to be useful appears to be the discriminant analysis. There are a number of reasons for this. First, the other indices are simpler and less costly to employ than the discriminant analysis, especially for a large number of groups. Second, forming a scale which best discriminates between groups might cause some items to be subtracted in constructing the scale (as occurred here). Besides being a not very appealing alternative to researchers and practitioners, interpretation problems arise. For example, on a scale with both positively and negatively weighted items, a student may have a lower score because he(she) answered all the items correctly. Thus, the discriminant analysis is ruled out as a viable alternative.

Of the two remaining techniques, the between-group item-total correlations appeared to be more appealing in this study. The intraclass correlations seem less useful because the more stringent cutoff does not lead to a scale more sensitive to between-group differences. In contrast, the between-class regression coefficients for three of the ALT variables (allocated time, engagement rate, and low error rate) increased when the top five items were selected on the basis of the between-class item-total correlations, as opposed to the top ten. However, we concede that this comparison may be study specific until the results of similar analyses on other data sets support this finding.

Finally, the stability of each of the indices of the between-group item discrimination indices was examined. By comparing the scales formed using the data from the BTES final study (Phase III-B) with the scales formed from pilot study data (Phase III-A Continuation), it was determined that none of the indices of item discrimination between groups were particularly stable across samples for the present data. However, the limited number of groups in the pilot study (5 classes) might be at least partially responsible for the observed instability.

Patterns of Item Response

In Chapter 5, patterns of correct and incorrect item responses were examined as a possible group phenomenon. It was determined that the patterns of response are indeed related to group membership. Using the pretest and the posttest patterns of response from the Phase III-B of the

BTES study, a number of conclusions were reached.

The patterns of correct item response on the posttest clearly showed a relationship to instructional coverage and emphasis that were not visible prior to instruction. In addition, the differences in coverage and emphasis are best described at the item level. That is, subscores of a test which can be logically foreseen do not always reflect the differences in coverage and emphasis which do exist. For example, the fractions subtest from the BTES study can logically be divided into four subscores (addition, subtraction, algebraic manipulation, and recognizing equivalent forms), but even at that level, differences in coverage and emphasis exist (e.g., see class 19 on the subtraction items).

However, the patterns of item response cannot be examined in a large-scale evaluation. The dimensions of the matrix (i.e., number of items and number of groups) in a large-scale evaluation can rapidly make the patterns of item response unmanageable. Consequently, an index of the patterns of item response is needed. For this study, Sato's (1980) caution index was used, and no differences between classes were found. However, we believe this to be a function of the limitations of the data set employed (see discussion below).

Besides analyzing patterns of correct item responses, patterns of incorrect responses were analyzed. Findings indicate that errors can occur in various ways. First, errors may be simply random (e.g., guessing or carelessness). Second, errors may be due to a systematic thought process. This systematic error can, in turn, be influenced by a number of circumstancs. Either the student may have adapted a faulty algorithm for problem solving independent of the classroom (e.g., absent on the day of instruction, or receives a faulty algorithm from someone outside of the class) or because of common experiences in the classroom. The reason for a common incorrect pro lem solving procedure in a classroom may be poor instruction resulting in in incorrect algorithm or instruction which has failed to correct a faulty procedure present prior to instruction. In this study, errors were shown to be a group phenomenon. Furthermore, systematic errors were found that were present both before and after instruction, as well as those that were not present before instruction, but were present after instruction.

Suggestions for Further Research

As with any research, the conclusions of this study are limited by the data employed. Consequently, a number of suggestions arise for further research.

Narrowness of Ability Range

Comparing the data from Phase III-B and Phase III-A Continuation of the BTES shows the effect of narrowing the ability range. By selecting six students from the middle of the overall distribution for each class, the within-class relationships are restricted in range. However, the between-class relationships may be more accurate (not affected by outliers). Furthermore, since the classroom was the unit of concern here, this did not seem like a bad restriction. In fact, the analyses of patterns of correct item response are probably helped by the restriction in ability range. Since no information about patterns of response is gained from a student who knows everything or nothing (in either case Sato's caution index equals zero), restricting the ability range may not drastically affect the results. In fact, more information about instructional coverage and emphasis may be available with the restricted ability range, since high ability students may get answers right, independent of the instruction, and low ability students may answer incorrectly, regardless of the instruction. However, further investigations without a restriction in the ability range seen warranted to test these assumptions.

Nature and Narrowness of Content

Fractions is the area of mathematics which receives the most attention in the fifth grade math curriculum for the BTES classes. The curriculum for fractions may be more uniform than in some other areas of mathematics instruction. This restricted range may have decreased the likelihood of finding large differences in instructional effects and perhaps limited the utility of the various item selection indices and the caution index.

The caution index, in particular, may play a more important role in tests with a broader coverage of topics. For example, if a test covered only two areas which are not always covered in the mathematics curriculum, then four different groups of classes could be identified at the subtest level (those that covered both areas, those that covered neither area, and those that covered one area or the other). We believe this to be the case for Harnisch and Linn (1981). They found school-to-school differences in the caution index. However, they used a more general test, which probably covered more content areas. While the content of the test was not discussed, we assume that a "mathematics" test would cover multiple content areas, some of which might be included in the curriculum of some schools and not in others.

Consequently, more research is needed to determine the effect of the range of the content covered.

The nature of mathematics makes it easier to discuss subsets of content cohesive items. However, these sorts of analyses may also be applicable in areas such as reading. Then groups of items might be examined from a different perspective (e.g., in terms of information processing or cognitive processes).

Grade Level

While this study focussed on the fifth grade, the curriculum at different grade levels may affect the item selection indices and the caution index. The typical curriculum tends to diversify as a function of grade level. For example, in the early grades everyone must learn their basic skills (e.g., addition, subtraction, multiplication). However, the higher grades often include material which is not taught to everyone (e.g., algebra, geometry). Consequently, these strategies for measuring differences in instructional coverage may be more distinctly defined in the later grades.

Concluding Remarks

In this study we have shown that information about differences between groups, especially in instruction, is available from an analysis of achievement test item data which is not available from the total test score. Items will vary to the extent that they reflect differences between groups. Furthermore, the patterns of item response, which reflect instructional coverage and emphasis, will vary from group-to-group.

While some of the hypothesized effects were not present in the BTES data base, further research is needed to check which effects were data specific and which effects were not found because of limitations of the data set. Assuming that some of the effects, which were absent in this study, might be found in future research, some conclusions about the utility of these techniques wil be drawn.

First, if the item selection strategies were found to be stable across samples when more groups are used, school effects studies and large-scale evaluations will be able to construct scales which are more sensitive to differences between groups, especially instructional and program variables. If the sca s are stable across samples, instruments could be formed a priori using the indices of item discrimination between groups rather than the traditional psychometric model used in test development.

Second, the patterns of item responses can be used in various contexts. They can be used in large-scale evaluations and school effects research to better measure what students know and how their knowledge compares with some relevant group (class, school, etc.). Furthermore, the patterns of responses can be used to discuss group-to-group differences in instructional coverage and emphasis. Finally, the patterns of item response can be used as an instructional adjunct. By measuring areas of coverage (or lack of coverage) and the errors in problem solving, suggestions for changes in the instructional sequence might appear. For example, does teaching abstract reasoning skills lead to certain types of problem solving skills and if so, is that the desired outcome? Or should certain incorrect algorithms ever by addressed or should only the correct algorithm be addressed?

The strategies outlined in this study and their utility still require further validation. However, in this study, we have shown that real and meaningful differences do exist between instructional groups that can only be tapped from achievement test item data (as opposed to test or subtest scores).

FOOTNOTES

[1] The item means on occasion B were .47, .44, .51, .40, .10, .44, .37, .09, .06, .27, .37, .21, .15, .32, and .20 for items 1 to 15, respectively. The item means on occasion C were .75, .73, .49, .61, .37, .67, .61, .28, .25, .54, .56, .47, .32, .53, and .36 for items 1 to 15, respectively.

[2] It is assumed with the computer packages available to social scientists (e.g., SPSS, BMDP, SAS) that the item-total correlation is simpler to get than the average correlation, which leads to the same results.

[3] Students with a no response on all fifteen items were excluded from this study to eliminate the effect of absentees. Thus, no response represents an alternative selected by the student.

APPENDIX A

FRACTION ITEMS IN FINAL STUDY

(PHASE III-B)

Subtraction of Fractions. Find the differences.

1	$\frac{5}{8} - \frac{3}{8} =$	2	$\frac{2}{16}$	$\frac{2}{8}$	$\frac{4}{5}$
2	$\frac{10}{12} - \frac{5}{12} =$	$\frac{5}{24}$	$\frac{5}{12}$	$\frac{7}{12}$	5
3	$8\frac{1}{3} - \frac{1}{3} =$	8	$7\frac{2}{3}$	7	$7\frac{1}{6}$
4	$8\frac{7}{9}$ $-\ 3\frac{2}{9}$	$5\frac{5}{18}$	4	$5\frac{1}{2}$	$5\frac{5}{9}$
5	$5\frac{5}{8}$ $-\ 4\frac{1}{4}$	$1\frac{3}{8}$	$1\frac{4}{8}$	2	$1\frac{4}{12}$

Addition of Fractions. Find the sums.

6	$\frac{3}{7} + \frac{2}{7} =$	$\frac{5}{7}$	$\frac{5}{14}$	$\frac{7}{14}$	$\frac{14}{21}$
7	$5\frac{1}{6} + \frac{4}{6} =$	$5\frac{5}{12}$	$5\frac{5}{6}$	$5\frac{6}{24}$	$5\frac{1}{2}$
8	$\frac{2}{3} + \frac{1}{3} =$	$\frac{3}{6}$	3	$\frac{2}{9}$	1
9	$\frac{3}{7} + \frac{5}{7} =$	$\frac{28}{35}$	$\frac{8}{14}$	$1\frac{1}{7}$	$\frac{4}{7}$
10	$9\frac{5}{8}$ $+ 6\frac{2}{8}$	$16\frac{7}{8}$	$15\frac{7}{8}$	$15\frac{16}{40}$	$15\frac{7}{16}$

Equivalent Fractions

11	$\frac{1}{2}$ equals:	$\frac{2}{3}$	$\frac{3}{6}$	$\frac{2}{5}$	$\frac{4}{9}$
12	$\frac{6}{8}$ equals:	$\frac{3}{9}$	$\frac{2}{6}$	$\frac{3}{4}$	$\frac{5}{7}$
13	$\frac{2}{3}$ equals:	$\frac{8}{12}$	$\frac{3}{4}$	$\frac{3}{9}$	$\frac{5}{6}$
14	What does N equal? $\frac{3}{8} = \frac{6}{N}$	12	11	16	24
15	What does N equal? $\frac{2}{7} = \frac{N}{21}$	3	16	7	6

APPENDIX B

FRACTION ITEMS IN PILOT STUDY

(PHASE III-A CONTINUATION)

1

$\frac{1}{2}$ is equivalent to which of these fractions?

A. $\frac{2}{3}$ B. $\frac{3}{6}$ C. $\frac{7}{15}$ D. $\frac{2}{5}$

2

$\frac{2}{3}$ is equal to?

A. $\frac{3}{9}$ B. $\frac{3}{4}$ C. $\frac{4}{10}$ D. $\frac{8}{12}$

3

$\frac{2}{5}$ is equal to?

A. $\frac{6}{15}$ B. $\frac{4}{12}$ C. $\frac{3}{7}$ D. $\frac{1}{3}$

4

What does N equal?

$\frac{3}{8} = \frac{6}{N}$

A. 12 . B. 24 C. 16 D. 11

5

What does N equal?

$\frac{2}{N} = \frac{10}{15}$

A. 3 B. 7 C. 8 D. 12

6

Find the value of N.

$\frac{2}{7} = \frac{N}{21}$

A. 5 B. 3 C. 6 D. 9

7

$\frac{6}{8}$ is equivalent to

A. $\frac{1}{3}$ B. $\frac{2}{6}$ C. $\frac{3}{4}$ D. $\frac{3}{9}$

8

$\frac{8}{24}$ is equivalent to

A. $\frac{2}{8}$ B. $\frac{1}{4}$ C. $\frac{4}{6}$ D. $\frac{1}{3}$

9

$\frac{5}{30}$ is equivalent to

A. $\frac{1}{25}$ B. $\frac{1}{6}$ C. 6 D. $\frac{1}{5}$

10

$\frac{12}{18}$ is equivalent to

A. $\frac{2}{4}$ B. $\frac{3}{4}$ C. $\frac{3}{8}$ D. $\frac{4}{6}$

11

$\frac{3}{7} + \frac{2}{7} = \square$

A. $\frac{1}{7}$ B. $\frac{5}{14}$ C. $\frac{5}{7}$ D. $\frac{7}{14}$

12

$\frac{2}{3} + \frac{1}{3} = \square$

A. $\frac{1}{3}$ B. $\frac{3}{6}$ C. 1 D. $\frac{3}{9}$

13

$\frac{2}{3} + \frac{1}{6} = \square$

A. $\frac{5}{6}$ B. $\frac{3}{9}$ C. $\frac{3}{6}$ D. $\frac{3}{18}$

14

$\frac{3}{5} + \frac{4}{10} = \square$

A. $\frac{9}{10}$ B. 1 C. $\frac{7}{15}$ D. $\frac{4}{5}$

15

$\frac{3}{7} + \frac{5}{7} = \square$

A. $\frac{8}{14}$ B. 1 C. $\frac{4}{7}$ D. $1\frac{1}{7}$

16

$\frac{5}{8} + \frac{6}{8} = \square$

A. $\frac{1}{8}$ B. $1\frac{1}{8}$ C. $\frac{11}{16}$ D. $1\frac{3}{8}$

17

$\frac{9}{10} + \frac{8}{10} = \square$

A. $1\frac{7}{10}$ B. $\frac{17}{20}$ C. $1\frac{3}{10}$ D. $\frac{1}{10}$

18

$5\frac{1}{6} + \frac{4}{6} = \square$

A. $5\frac{1}{2}$ B. $5\frac{5}{6}$ C. $5\frac{5}{12}$ D. $6\frac{5}{6}$

19

$$\begin{array}{r} 9\frac{5}{8} \\ +\ 6\frac{2}{8} \\ \hline \end{array}$$

A. $15\frac{3}{16}$ C. $15\frac{7}{8}$

B. $16\frac{7}{8}$ D. $15\frac{7}{16}$

20

$$\begin{array}{r} 8\frac{1}{2} \\ +\ 2\frac{3}{6} \\ \hline \end{array}$$

A. 11 C. $10\frac{7}{8}$

B. $10\frac{1}{2}$ D. 10

21

$\frac{5}{8} - \frac{3}{8} =$ □

A. $\frac{1}{8}$ B. 2 C. $\frac{2}{16}$ D. $\frac{2}{8}$

22

$13 - \frac{3}{8} =$ □

A. $12\frac{1}{4}$ B. $13\frac{5}{8}$ C. $12\frac{5}{8}$ D. $12\frac{7}{8}$

23

$5 - \frac{1}{4} =$ □

A. $4\frac{3}{4}$ B. $5\frac{1}{4}$ C. 4 D. $3\frac{3}{4}$

24

$\frac{10}{12} - \frac{5}{12} =$ □

A. $\frac{2}{7}$ B. $\frac{7}{12}$ C. $\frac{5}{24}$ D. $\frac{5}{12}$

25

$16 - \frac{3}{9} =$ □

A. $15\frac{1}{9}$ B. $15\frac{6}{9}$ C. $14\frac{9}{12}$ D. $16\frac{6}{9}$

26

$8\frac{1}{3} - \frac{1}{3} =$ □

A. $7\frac{1}{3}$ B. $7\frac{2}{6}$ C. $8\frac{1}{9}$ D. 8

27

$$\begin{array}{r} 5\frac{5}{8} \\ -\ 4\frac{4}{8} \\ \hline \end{array}$$

A. $1\frac{1}{8}$ B. $1\frac{1}{16}$

C. $1\frac{1}{4}$ D. $\frac{1}{8}$

28

$7\frac{1}{2} - \frac{2}{5} =$ □

A. $6\frac{1}{10}$ B. $7\frac{1}{10}$ C. $6\frac{3}{7}$ D. $7\frac{1}{5}$

29

$8\frac{7}{9} - 3\frac{2}{9} =$ □

A. $4\frac{5}{9}$ B. $5\frac{5}{9}$ C. $5\frac{5}{18}$ D. $4\frac{10}{18}$

30

$$\begin{array}{r} 9\frac{5}{6} \\ -\ 2\frac{5}{12} \\ \hline \end{array}$$

A. $7\frac{5}{18}$ C. $7\frac{5}{12}$

B. $6\frac{5}{12}$ D. 6

APPENDIX C

ZERO ORDER CORRELATIONS OF CONSTRUCTED SCALES WITH PRETEST, ALLOCATED TIME (A.T.), ENGAGEMENT RATE (ENG.), LOW ERROR RATE (L.E.R.), AND HIGH ERROR RATE (H.E.R.)

1. Correlations of scales formed from intraclass correlations

	Top 5	Top 10
Class Level		
PRETEST	.30	.41
A.T.	.40	.45
ENG.	.20	.22
L.E.R.	-.32	-.29
H.E.R.	-.16	-.19
Individual Level		
PRETEST	.33	.47
A.T.	.38	.43
ENG.	.25	.25
L.E.R.	-.07	-.05
H.E.R.	-.27	-.26

2. Correlations of scales formed from average between-class item correlations

	Top 5	Top 10
Class Level		
PRETEST	.47	.46
A.T.	.40	.38
ENG.	.21	.17
L.E.R.	-.27	-.27
H.E.R.	-.17	-.21
Individual Level		
PRETEST	.57	.54
A.T.	.36	.38
ENG.	.22	.18
L.E.R.	-.05	-.06
H.E.R.	-.25	-.25

3. Correlations of scales formed from discriminant analyses

	Discriminant Function			
	1	2	3	4
Class Level				
PRETEST	.30	.36	.08	.25
A.T.	.01	.27	.16	-.22
ENG.	-.16	.12	-.22	-.06
L.E.R.	-.05	-.37	-.12	.00
H.E.R.	-.16	-.18	.09	-.19
Individual Level				
PRETEST	.19	.34	-.04	.12
A.T.	.04	.31	.11	-.17
ENG.	-.04	.15	-.07	-.13
L.E.R.	.06	-.20	-.02	-.12
H.E.R.	-.03	-.19	.05	-.06

4. Correlations of scales formed from between-class correlations of items with allocated time

	Top 5	Top 10
Class Level		
PRETEST	.48	.50
A.T.	.51	.48
ENG.	.22	.22
L.E.R.	-.24	-.26
H.E.R.	-.12	-.15
Individual Level		
PRESTEST	.50	.53
A.T.	.47	.45
ENG.	.28	.25
L.E.R.	.02	-.01
H.E.R.	-.22	-.21

APPENDIX D

STUDENT PATTERNS OF ITEM RESPONSE,
SATO'S (1980) CAUTION INDEX (C)
AND
HARNISCH AND LINN'S (1981)
MODIFIED CAUTION INDEX (C*)

```
*JOB
      INTEGER TEST(100,300),ANSWER(100),FTEST(300)
      REAL INDEX(2)
      CHARACTER*9 ID(300)
      DIMENSION ITEMD(300),ITOT(300),SI(120),PI(120),SJ(120),PJ(120)
      COMMON /SPINFO/TEST,ANSWER,TOT,K,L,FTEST,T,ID,INDEX
      DIMENSION IFMT(20),IOS(300),IOP(100),TIND(2,300)
      DIMENSION SJ1(100),PJ1(100),IN(100),ITMP(3,100)
      DIMENSION DT1(100),DT2(100),IND(100)
      CHARACTER*1 SI,PI,SJ1,PJ1,COLON,BLANK,BAR,DASH1,DOT1,DT1,IFMT*4
      CHARACTER*2 SJ,PJ,BLANK2,DOT,DASH,DT2
      DATA BLANK/' '/,COLON/':'/,BAR/'|'/,DOT1/'.'/,DASH1/'-'/
      DATA BLANK2/'  '/,DOT/'..'/,DASH/'--'/
      DATA DT1/100*'.'/,DT2/100*' .'/
C          THIS PROGRAM WAS WRITTEN BY CHIH-PING CHOU
C              WITH ASSISTANCE FROM DAVID MCARTHUR,
C             FOLLOWING SATO (1980), AND COLLEAGUES.
C              CENTER FOR STUDY OF EVALUATION, UCLA
C                        SEPTEMBER 1981
C ****TO BEGIN, READ IN 3 CARDS WHICH PROVIDE THE FOLLOWING INFORMATION :
C * 1.    N OF CASES  AND  N OF ITEMS BY THE FORMAT OF (I4,I3)    ****
C * 2.    CORRECT ANSWERS FOR EACH ITEM BY THE FORMAT OF (F1.0)****
C * 3.    THE FORMAT FOR SUBJECT INFORMATION (ID OF UP TO 4 CHARACTERS
C            FOLLOWED BY UP TO 100 SCORES), FOR EXAMPLE: (A4,6X,50F1.0)
C        THEN GIVE THIS CARD FOR LOCATION OF DATASET:
C   //FT08F001 DD DSN=AAAAAII.NAME,DISP=OLD
C
      READ(5,10) NSUB,NITEM,(ANSWER(I),I=1,NITEM)
   10 FORMAT(I4,I3/80I1)
      READ(5,20) (IFMT(I),I=1,20)
   20 FORMAT(20A4)
C ----------        READ IN SUBJECT'S INFORMATION    -------------
      DO 30 I=1,NITEM
      IN(I)=I
   30 ITOT(I)=0
      IT=0
      DO 50 K=1,NSUB
      READ(8,IFMT) ID(K),(TEST(I,K),I=1,NITEM)
      TOT=0.0
      CALL SCORE(NITEM)
      DO 40 I=1,NITEM
      ITOT(I)=ITOT(I)+TEST(I,K)
   40 TOT=TOT+TEST(I,K)
      ITEMD(K)=TOT
   50 CONTINUE
      PRINT 60
   60 FORMAT(1H1)
      T=FLOAT(NITEM)/10.0+0.95
      NTM=T
      MITM=3
      IF(NITEM .LT. 100) MITM=2
      IF(NITEM .LT. 10) MITM=1
      MSUB=4
      IF(NSUB .LT. 1000) MSUB=3
      IF(NSUB .LT. 100) MSUB=2
      IF(NSUB .LT. 10)  MSUB=1
      CALL ORDER(ITEMD,IOS,NSUB)
      CALL ORDER(ITOT,IOP,NITEM)
      PRINT 100
  100 FORMAT (' ITEMS, IN ASCENDING ORDER OF DIFFICULTY')
```

```
      CALL VERWRT(IN,NITEM,NSUB,ITMP)
      IF(NITEM .GT. 50) GO TO 1000
      DO 102 I=1,NITM
  102 WRITE(6,90) (ITMP(I,J),J=1,NITEM)
   90 FORMAT(12X,50I2)
   91 FORMAT(12X,100I1)
      PRINT 95
   95 FORMAT(//)
      CALL VERWRT(IOP,NITEM,NITM,ITMP)
      DO 103 I=1,NITM
  103 WRITE (6,90) (ITMP(I,J),J=1,NITEM)
      PRINT 95
      PRINT 110
  110 FORMAT (' SUBJECTS, IN'/' DESCENDING ORDER:'/' RANK|ID',T115,
     1' TOTAL;C.S   C.S*')
      T=0.0
      DO 70 I=1,NITEM
      K=NITEM+1-I
      IF(ITOT(K) .EQ. 0) GO TO 70
      IP=2*K+1
      PI(IP)=COLON
      JP=K
      GO TO 80
   70 CONTINUE
   80 DO 120 J=1,NITEM
  120 T=T+ITOT(J)
      T=T/FLOAT(NITEM)
      NK=NITEM
      NCOUNT=0
      DO 130 I=1,120
      SI(I)=BLANK
      PI(I)=BLANK
      SJ(I)=BLANK2
  130 PJ(I)=BLANK2
      DO 260 K=1,NSUB
      L=IOS(K)
      TOT=ITEND(K)
      DO 140 I=1,NITEM
      I1=IOP(I)
  140 PTEST(I)=TEST(I1,L)
      JS=ITEND(K)
      IF((K+1) .GT. NSUB) GO TO 160
      JS1=ITEND(K+1)+1
      IF(JS .LT. JS1 .OR. JS .EQ. 0) GO TO 160
      DO 150 I=JS1,JS
  150 SJ(I)=DASH
  160 IS=ITEND(K)*2+1
      SI(IS)=BAR
      NP=NK
      I1=0
      DO 170 I=1,NITEM
      IF(K .GT. ITOT(I)) GO TO 180
      IF(K .NE. ITOT(I)) GO TO 170
      I1=I1+1
      IF(I1 .EQ. 1)NK=I-1
      NP=I
      PJ(NP)=DOT
  170 CONTINUE
  180 JP=NP
      IP=JP*2+1
      PI(IP)=COLON
      NKP=ITEND(K)-NP
      NKP=IABS(NKP)
      NCOUNT=NCOUNT+NKP
      PRINT 190,(SI(I),I=1,IS)
  190 FORMAT(12X,120A1)
      PRINT 200,(PI(I),I=1,IP)
  200 FORMAT('+',11X,120A1)
      CALL CAUTN2(NITEM,NSUB,ITOT)
      CTEST=INDEX(1)
      DTEST=INDEX(2)
      IT=IT+ITEND(K)
      WRITE(6,210)K,ID(L),(FTEST(I),I=1,NITEM)
      GO TO (212,212,213,214,215,213,214,214,215,215),NTM
  212 WRITE(6,222)ITEND(K),CTEST,DTEST
      GO TO 218
  213 WRITE(6,223)ITEND(K),CTEST,DTEST
      GO TO 218
  214 WRITE(6,224)ITEND(K),CTEST,DTEST
      GO TO 218
  215 WRITE(6,225)ITEND(K),CTEST,DTEST
  218 IF (JS .EQ. 0) GO TO 240
      PRINT 230,(SJ(I),I=1,JS)
  230 FORMAT(12X,60A2)
  240 IF (JP .EQ. 0) GO TO 260
      PRINT 250,(PJ(I),I=1,JP)
  250 FORMAT('+',11X,60A2)
  222 FORMAT('+',15X,40X,I3,2F6.3)
  223 FORMAT('+',15X,60X,I3,2F6.3)
  224 FORMAT('+',15X,80X,I3,2F6.3)
  225 FORMAT('+',15X,100X,I3,2F6.3)
  210 FORMAT('+',I5,A5,1X,51I2)
  260 CONTINUE
      WRITE(6,270) NCOUNT
  270 FORMAT(/80X,'COUNT OF CELLS BETWEEN S & P CURVE :',I5)
      PPINT 280
  280 FORMAT (//' ITEM TOTALS: ')
      CALL VERWRT(ITOT,NITEM,NSUB,ITMP)
      DO 285 I=1,NSUB
  285 WRITE(6,90) (ITMP(I,J),J=1,NITEM)
      GO TO 2000
 1000 DO 1102 I=1,NITM
 1102 WRITE(6,91) (ITMP(I,J),J=1,NITEM)
      PRINT 95
      CALL VERWRT(IOP,NITEM,NITM,ITMP)
      DO 1103 I=1,NITM
 1103 WRITE (6,91) (ITMP(I,J),J=1,NITEM)
      PRINT 95
      PRINT 110
      T=0.0
      DO 1120 J=1,NITEM
 1120 T=T+ITOT(J)
      T=T/FLOAT(NITEM)
      NK=NITEM
      NCOUNT=0
      DO 1130 I=1,100
      SJ1(I)=BLANK
 1130 PJ1(I)=BLANK
      DO 1260 K=1,NSUB
      L=IOS(K)
      TOT=ITEMD(K)
      DO 1140 I=1,NITEM
      I1=IOP(I)
 1140 FTEST(I)=TEST(I1,L)
      JS=ITEMD(K)
      IF((K+1) .GT. NSUB) GO TO 1160
      JS1=ITEMD(K+1)+1
      IF(JS .LT. JS1 .OR. JS .EQ. 0) GO TO 1160
      DO 1150 I=JS1,JS
 1150 SJ1(I)=DA[illegible]1
 1160 IS=ITEMD(K)*2+1
      NP=NK
      I1=0
      DO 1170 I=1,NITEM
      IF(K .GT. ITOT(I)) GO TO 1180
      IF(K .NE. ITOT(I)) GO TO 1170
      I1=I1+1
      IF(I1 .EQ. 1)NK=I-1
      NP=I
      PJ1(NP)=DOT1
 1170 CONTINUE
 1180 JP=NP
      IP=JP*2+1
      NKP=ITEMD(K)-NP
      NKP=IABS(NKP)
      NCOUNT=NCOUNT+NKP
      CALL CAUTN2(NITEM,NSUB,ITOT)
      CTEST=INDEX(1)
      DTEST=INDEX(2)
      IT=IT+ITEMD(K)
      WRITE(6,211)K,ID(L),(FTEST(I),I=1,NITEM)
      GO TO (1212,1212,1213,1214,1215,1213,1214,1214,1215,1215),NTM
 1212 WRITE(6,222)ITEMD(K),CTEST,DTEST
      GO TO 1218
 1213 WRITE(6,223)ITEMD(K),CTEST,DTEST
      GO TO 1218
 1214 WRITE(6,224)ITEMD(K),CTEST,DTEST
      GO TO 1218
 1215 WRITE(6,225)ITEMD(K),CTEST,DTEST
 1218 IF (JS .EQ. 0) GO TO 1240
      PRINT 231,(SJ1(I),I=1,JS)
  231 FORMAT(12X,100A1)
 1240 IF (JP .EQ. 0) GO TO 1260
      PRINT 251,(PJ1(I),I=1,JP)
  251 FORMAT('+',11X,100A1)
  211 FORMAT(I6,A5,1X,100I1)
 1260 CONTINUE
      WRITE(6,270)NCOUNT
      PRINT 280
      CALL VERWST(ITOT,NITEM,NSUB,ITMP)
      DO 1285 I=1,NSUB
 1285 WRITE(6,91)(ITMP(I,J),J=1,NITEM)
C   REVERSE THE TEST MATRIX FOR THE INDEX COEFFICIENTS OF ITEM
 2000 T=T*NITEM/FLOAT(NSUB)
      DO 510 K=1,NITEM
      L=IOP(K)
      TOT=ITOT(K)
      DO 300 I=1,NSUB
      I1=IOS(I)
  300 FTEST(I)=TEST(L,I1)
```

```
      CALL CAUTN2(NSUB,NITEM,ITEMD)
      TIND(1,K)=INDEX(1)
  510 TIND(2,K)=INDEX(2)
      PRINT 320
  320 FORMAT (//,' CAUTION INDICES FOR ITEMS:')
      DO 330 L=1,2
      DO 345 J=1,NITEM
  345 IND(J)=TIND(L,J)*100.0+0.5
      CALL VERWRT(IND,NITEM,3,ITMP)
      IF(NITEM .GT. 50) GO TO 3000
      DO 340 I=1,3
      IF(I .NE. 2) GO TO 360
      WRITE(6,350) (DT2(J),J=1,NITEM)
  350 FORMAT(12X,50A2)
  351 FORMAT(12X,100A1)
  360 WRITE(6,90) (ITMP(I,J),J=1,NITEM)
  340 CONTINUE
      GO TO 330
 3000 DO 3340 I=1,3
      IF(I .NE. 2) GO TO 3360
      WRITE(6,351) (DT1(J),J=1,NITEM)
 3360 WRITE(6,91) (ITMP(I,J),J=1,NITEM)
 3340 CONTINUE
      PRINT 95
  330 CONTINUE
      DO 430 J=1,NITEM
      PRINT 420,J,IOP(J),ITOT(J),(TIND(I,J),I=1,2)
  420 FORMAT (' ITEM RANK = ',I4,'  ITEM # = ',I4,'  ITEM TOTAL = ',
     1I5,'  C = ',F5.2,'   C* = ',F5.2)
  430 CONTINUE
  440 T=NSUB*NITEM
      PM=SQRT(T)+.5
      M=PM
      DM=.420
      IF(M .GT. 95) GO TO 450
      DM=.25477957+.00376604*M-.00002202*M**2
  450 P=FLOAT(IT)/FLOAT(NSUB)/FLOAT(NITEM)
      DSTAR=FLOAT(NCOUNT)/(4*NSUB*NITEM*P*(1-P)*DM)
      PRINT 460,P,DSTAR
  460 FORMAT(//50X,'  P  =',F8.3,10X,'  D*  =',F8.3)
      PRINT 100
      STOP
      END
C ****
C-----------------------------
C ****

      SUBROUTINE SCORE(NITEM)
      INTEGER TEST(100,300),ANSWER(100),FTEST(300)
      REAL INDEX(2)
      CHARACTER*9 ID(300)
      COMMON /SPINFO/TEST,ANSWER,TOP,K,L,FTEST,T,ID,INDEX
      DO 10 I=1,NITEM
      IF(TEST(I,K) .NE. ANSWER(I)) TEST(I,K)=0
   10 IF(TEST(I,K) .EQ. ANSWER(I)) TEST(I,K)=1
      RETURN
      END
C ********
C -----------------------------
C ********
```

```
      SUBROUTINE VEFWRT(IA,N,M,IT)
      DIMENSION IA(100),IT(3,100)
      DO 1 I=1,M
      DO 1 J=1,N
      IT(I,J)=FLOAT(IA(J))/10.0**(M-I)
      DO 2 K=1,I
      K1=I-K
      IF(K1 .EQ. 0) GO TO 1
      IT(I,J)=IT(I,J)-IT(K,J)*10**(K1)
    2 CONTINUE
    1 CONTINUE
      RETURN
      END
C *******
C------------------------------
C *******

      SUBROUTINE ORDER(IT,IO,N)
      DIMENSION IT(N),IO(N)
      N1=N-1
      DO 10 I=1,N
   10 IO(I)=I
      DO 20 I=1,N1
      M=N-I
      DO 20 J=1,M
      J1=J+1
      IF(IT(J) .GE. IT(J1)) GO TO 20
      TEMP=IT(J1)
      IT(J1)=IT(J)
      IT(J)=TEMP
      TEMP=IO(J1)
      IO(J1)=IO(J)
      IO(J)=TEMP
   20 CONTINUE
      RETURN
      END
C ****
C------ ----------------------
C ****

      SUBROUTINE CAUTN2(NITEM,NSUB,ITOT)
      INTEGER TEST(100,300),ANSWER(100),FTEST(300)
      REAL INDEX(2)
      CHARACTER*9 ID(300)
      DIMENSION ITOT(300)
      COMMON /SPINFO/TEST,ANSWER,TOT,K,L,FTEST,T,ID,INDEX
C*
C  CALCULATE CAUTION INDEX : INDEX(1) & MODIFIED CAUTION INDEX : INDEX(2)  *
C*
      TITM=FLOAT(NITEM)
      IF(TOT .EQ. 0.0) GO TO 60
      IF(TOT .EQ. TITM) GO TO 60
      T1=0.0
      T2=0.0
      T3=0.0
      NSBJT=IFIX(TOT)
      NSBJT1=NSBJT+1
      DO 10 J=1,NSBJT
      T1=T1+(1.-FTEST(J))*ITOT(J)
   10 T2=T2+ITOT(J)
```

```
      IF(NSBJT1 .GT. NITEM) GO TO 30
      DO 20 J=NSBJT1,NITEM
   20 T3=T3+FTEST(J)*ITOT(J)
   30 INDEX(1)=(T1-T3)/(T2-TOT*T)
      T4=0.0
      JMNSP=NITEM+1-NSBJT
      IF(JMNSP .GT. NITEM) GO TO 50
      DO 40 J=JMNSP,NITEM
   40 T4=T4+ITOT(J)
   50 INDEX(2)=(T1-T3)/(T2-T4)
      GO TO 70
   60 INDEX(1)=0.0
      INDEX(2)=0.0
   70 CONTINUE
      RETURN
      END

    *RUN
```

```
ITEMS, IN ASCENDING ORDER OF DIFFICULTY
                 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1
                 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5


                 0 0 0 0 0 1 1 0 1 1 0 1 1 0 0
                 1 2 6 4 7 1 0 3 4 2 5 5 3 8 9


SUBJECTS, IN
DESCENDING ORDER;
RANK|CLASS                                                   TOTAL  C     C*
    111          1 1 1 1 1 1 1 1 1 1 1 1 1 1 1|                 15 0.000 0.000

    211          1 1 1 1 1 1 1 1 1 1 1 1 1 1 1|                 15 0.000 0.000

    323          1 1 1 1 1 1 1 1 1 1 1 1 1 1 1|                 15 0.000 0.000

    424          1 1 1 1 1 1 1 1 1 1 1 1 1 1 1|                 15 0.000 0.000

    527          1 1 1 1 1 1 1 1 1 1 1 1 1 1 1|                 15 0.000 0.000

    627          1 1 1 1 1 1 1 1 1 1 1 1 1 1 1|                 15 0.000 0.000

    727          1 1 1 1 1 1 1 1 1 1 1 1 1 1 1|                 15 0.000 0.000
                                             --
    8 4          1 1 1 1 1 1 1 1 1 1 1 0 1 1|1:                 14 0.307 0.149

    9 9          1 1 1 1 1 1 1 0 1 1 1 1 1 1|1:                 14 1.074 0.522

   1011          1 1 1 1 1 1 1 1 1 1 1 1 1 0|1:                 14 0.061 0.030

   1112          1 1 1 1 1 1 1 0 1 1 1 1 1 1|1·                 14 1.074 0.522

   1212          1 0 1 1 1 1 1 1 1 1 1 1 1 1|1:                 14 1.963 0.955

   1326          1 1 1 1 1 1 1 0 1 1 1 1 1 1|1:                 14 1.074 0.522
                                           --
   14 1          1 1 1 1 1 1 1 1 1 1 0 1 1|1 0:                 13 0.174 0.085

   15 3          1 1 1 0 1 1 1 1 1 1 1 1 1|0 1:                 13 0.807 0.395

   16 6          1 1 1 1 1 1 1 1 1 1 0 1 0|1 1:                 13 0.316 0.155

   1712          1 0 1 1 1 1 1 1 1 1 1 0 1|1 1:                 13 1.139 0.558

   1823          1 1 1 1 1 1 1 0 1 1 1 1 0|1 1:                 13 0.665 0.326

   1924          1 1 1 1 1 1 1 1 1 0 1 1 0|1 1:                 13 0.554 0.271

   2024          1 1 1 1 1 1 1 0 1 1 0 1 1|1 1:                 13 0.728 0.357
                                         --
   21 1          1 1 1 1 1 1 0 1 1 1 1 0|1 1 0:                 12 0.403 0.200

   22 3          1 1 1 1 1 1 1 1 1 0 0 1|1 1 0:                 12 0.346 0.171

   2311          1 1 1 0 1 1 1 1 1 1 1 0|0 1 1:                 12 0.680 0.337
```

```
2423     1 1 1 1 1 1 1 1 1 0 0 1|0 1 1:          12 0.449 0.223
2527     1 1 1 1 1 1 1 1 0 1 1 0|0 1 1:          12 0.495 0.246
26 8     1 1 1 1 1 1 0 1 1 1 0|1 1 0 0:          11 0.274 0.139
2710     1 1 1 1 1 1 1 1 0 1 1|1 0 0 0:          11 0.229 0.116
2818     1 1 1 1 0   1 1 1 1 0|1 1 0 0:          11 0.366 0.165
29 1     1 1 1 1 1 0 1 1 0 1|0 0 0 1 1:          10 0.550 0.285
30 8     1 1 1 1 1 1 1 1 0 1|0 1 0 0 0:          10 0.194 0.100
31 9     1 0 0 1 1 1 1 0 1 1|1 0 1 1:0           10 1.008 0.522
3211     1 1 1 1 1 1 1 1 0 1|0 0 1 0:0           10 0.202 0.104
3312     1 1 1 0 1 1 1 1 1 1|0 0 0:1 0           10 0.380 0.197
3421     1 1 1 1 1 1 0 1 1 1|0 1 0:0 0           10 0.202 0.104
3521     1 1 1 1 1 0 0 1 1 1|0 0 0:1 1           10 0.558 0.289
3624     1 1 1 1 1 1 1 1 1 0|0 0 0:1 0           10 0.202 0.104
3725     1 1 1 1 1 1 1 1 1 1|0 0 0:0 0           10 0.090 0,000
3827     1 1 1.1 1 0 1 0 1 0|1 0 0:1 1           10 0.667 0.345
39 1     1 1 1 1 1 1 0 0 0|1 0 0 1:1 0            9 0.501 0.259
40 5     1 1 1 1 1 1 1 1 0|0 0 0:0 0 1            9 0.262 0.135
41 5     1 1 1 1 1 1 1 1 0|0 0:0 0 0 1            9 0.262 0.135
42 8     1 1 0 1 1 1 0 1 1|1 1:0 0 0 0            9 0.374 0.193
4312     1 1 1 1 1 1 1 1 1|0 0:0 0 0 0            9 0.000 0.000
4417     1 1 1 1 1 1 1 1 0|0:0 0 1 0 0            9 0.195 0.100
4518     1 1 0 1 0 1 0 1 1|1:1 0 1 0 0            9 0.651 0.336
4618     1 1 0 1 0 1 0 0 1|1:1 1 1 0 0            9 0.838 0.432
4723     1 0 0 1 0 1 1 1 1|1:0 1 1 0 0            9 0.883 0.456
4823     1 1 0 1 1 1 1 0 1|1:0 1 0 0 0            9 0.389 0.201
4924     1 1 1 0 1 1 1 0 0|1:0 1 1 0 0            9 0.554 0.286
50 5     1 1 1 0 1 1 1 1|0 0:1 0 0 0 0            8 0.290 0.146
51 6     1 1 1 1 0 1 0 0|1 1:0 0 1 0 0            8 0.343 0.173
52 6     1 1 1 1 1 0 1 1|1 0:0 0 0 0 0            8 0.023 0.012
53 8     1 1 0 1 0 1 0 1|1 1:1 0 0 0 0            8 0.465 0.235
```

```
54 9     1 1 1 1 0 1 0 0|1 1:0 1 0 0 0          8 0.335 0.169
5510     1 1 1 1 1 0 1 1|0 0:1 0 0 0 0          8 0.191 0.096
5610     1 1 1 0 1 1 0 1|0 1:0 0 1 0 0          8 0.381 0.192
5711     1 1 0 1 0 1 0 0|1 1:1 0 1 0 0          8 0.663 0.335
5812     1 1 1 1 1 0 1 0|1 0:0 1 0 0 0          8 0.213 0.108
                            ••
5914     1 1 1 1 1 1 1 1|0:0 0 0 0 0 0          8 0.000 0.000
6019     1 1 1 1 1 1 0 1|0:1 0 0 0 0 0          8 0.061 0.031
6119     1 1 1 1 1 0 1 1|0:1 0 0 0 0 0          8 0.076 0.038
6221     1 1 1 1 1 0 0 0|1:1 0 1 0 0 0          8 0.274 0.138
6323     1 1 1 1 0 0 1 0|1:1 0 1 0 0 0          8 0.351 0.177
6427     1 1 1 1 1 0 0 1|0:0 1 0 0 0 1          8 0.465 0.235
                       --
65 1     1 1 1 0 1 0 0|1 0:1 0 0 0 1 0          7 0.466 0.231
                       ••••
66 4     1 1 1 1 0 1 0|0 1 0 0 1 0 0 0          7 0.287 0.142
                     ••
67 5     1 1 1 1 1 1:1|0 0 0 0 0 0 0 0          7 0.000 0.000
68 8     1 1 1 1 1 1:0|0 0 0 1 0 0 0 0          7 0.179 0.088
                   ••
6910     1 1 1 1 0:1 1|0 0 0 0 0 1 0 0          7 0.287 0.142
7014     1 1 1 1 1:0 0|1 1 0 0 0 0 0 0          7 0.031 0.015
7118     1 1 0 1 0:0 0|0 1 1 0 1 1 0 0          7 0.722 0.358
7221     1 1 1 1 1:0 0|1 1 0 0 0 0 0 0          7 0.031 0.015
7321     1 1 1 0 1:1 1|1 0 0 0 0 0 0 0          7 0.124 0.062
                   --
74 5     1 1 1 0 0:1|0 0 0 1 1 0 0 0 0          6 0.447 0.216
75 5     1 1 1 0 1:0|1 0 0 0 0 1 0 0 0          6 0.343 0.166
76 6     1 1 0 1 0:1|0 0 0 1 0 0 1 0 0          6 0.510 0.247
               ••
7710     1 1 0 0:0 0|1 0 1 0 1 0 1 0 0          6 0.774 0.375
7814     0 0 0 1:0 1|0 1 0 0 1 0 0 1 1          6 1.451 0.703
7914     1 1 1 1:1 0|1 0 0 0 0 0 0 0 0          6 0.016 0.008
8016     0 0 1 1:1 1|0 1 1 0 0 0 0 0 0          6 0.486 0.236
8117     1 1 1 1:1 0|1 0 0 0 0 0 0 0 0          6 0.016 0.008
             ••
8217     1 1 1:1 1 0|1 0 0 0 0 0 0 0 0          6 0.016 0.008
8324     1 1 1:0 1 0|1 0 0 0 1 0 0 0 0          6 0.319 0.154
```

```
8425     1 1 0:1 0 0|1 1 0 0 1 0 0 0 0                    6 0.439 0.212
8526     1 1 0:1 0 0|0 1 1 0 0 1 0 0 0                    6 0.470 0.228
86 6     1 1:1 1 0|0 0 1 0 0 0 0 0 0 0                    5 0.092 0.044
87 9     0 0:1 1 1|0 1 0 1 0 0 0 0 0 0                    5 0.500 0.241
8817     1 1:1 0 1|0 1 0 0 0 0 0 0 0 0                    5 0.125 0.060
8918     0 1:0 1 0|0 0 1 1 0 1 0 0 0 0                    5 0.708 0.341
9019     0 0:1 1 1|0 1 1 0 0 0 0 0 0 0                    5 0.500 0.241
9121     1 1:1 0 0|0 0 0 0 0 1 0 1 0 0                    5 0.625 0.301
9226     1 1:1 0 1|0 0 0 0 0 0 0 1 0 0                    5 0.350 0.169
93 3     0 0:0 0|0 1 0 0 1 1 1 0 0 0 0                    4 1.154 0.569
94 4     1 1:1 0|1 0 0 0 0 0 0 0 0 0 0                    4 0.047 0.023
9514     0:0 1 0|1 0 1 0 0 1 0 0 0 0 0                    4 0.675 0.333
9616      1:1 1 0|0 0 0 0 1 0 0 0 0 0 0                   4 0.150 0.074
9717      1:1 0 1|0 0 0 0 1 0 0 0 0 0 0                   4 0.188 0.093
9818     :0 1 0 0|0 1 0 0 1 1 0 0 0 0 0                   4 0.675 0.333
9919     :0 0 0 0|0 1 0 1 0 1 1 0 0 0 0                   4 1.154 0.569
10019    :0 0 1 1|1 0 0 1 0 0 0 0 0 0 0                   4 0.469 0.231
10121    :1 1 0 1|0 0 0 0 1 0 0 0 0 0 0                   4 0.188 0.093
10226    :1 1 1 0|0 1 0 0 0 0 0 0 0 0 0                   4 0.122 0.060
103 9    :1 0 1|0 0 0 1 0 0 0 0 0 0 0 0                   3 0.317 0.160
10416    :0 0 0|0 0 1 0 0 1 0 0 1 0 0 0                   3 1.168 0.589
10519    :0 0 0|0 1 0 0 1 0 1 0 0 0 0 0                   3 0.873 0.440
10625    :1 1 0|0 0 0 0 0 1 0 0 0 0 0 0                   3 0.227 0.114
10725    :0 0 0|0 1 0 1 0 0 0 1 0 0 0 0                   3 1.032 0.520
10825    :1 1 1|0 0 0 0 0 0 0 0 0 0 0 0                   3 0.000 0.000
10926    :1 1 1|0 0 0 0 0 0 0 0 0 0 0 0                   3 0.000 0.000
110 4    :0 1|0 0 0 0 1 0 0 0 0 0 0 0 0                   2 0.471 0.240
111 4    :0 0|0 1 0 0 0 1 0 0 0 0 0 0 0                   2 0.684 0.349
112 4    :1 0|0 0 0 0 1 0 0 0 0 0 0 0 0                   2 0.426 0.217
11310    :0 0|0 0 0 0 0 0 0 0 1 0 0 0 1                   2 1.793 0.915
```

```
11416     :0 0|0 0 0 0 0 0 1 0 0 1 0 0 0                    2 1.307 0.667

11516     :0 0|0 0 0 0 0 0 1 0 0 1 0 0 0                    2 1.307 0.667

11621     :0 0|0 0 0 0 0 1 0 0 0 0 1 0 0                    2 1.322 0.674
             --
11714     :0|0 0 0 0 0 0 1 0 0 0 0 0 0 0                    1 0.930 0.478

11821     :0|0 0 0 0 0 0 1 0 0 0 0 0 0 0                    1 0.930 0.478
           --
119 3     |0 0 0 0 0 0 0 0 0 0 0 0 0 0 0                    0 0.000 0.000
120 3     |0 0 0 0 0 0 0 0 0 0 0 0 0 0 0                    0 0.000 0.000
121 3     |0 0 0 0 0 0 0 0 0 0 0 0 0 0 0                    0 0.000 0.000
12216     |0 0 0 0 0 0 0 0 0 0 0 0 0 0 0                    0 0.000 0.000
12317     |0 0 0 0 0 0 0 0 0 0 0 0 0 0 0                    0 0.000 0.000
                              COUNT OF CELLS BETWEEN S & P CURVE :  256


ITEM TOTALS:
              0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
              9 9 8 8 7 6 6 6 6 5 4 4 3 3 3
              7 4 5 1 6 8 6 5 5 8 3 0 9 2 0


CAUTION INDICES FOR ITEMS:
              0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
              . . . . . . . . . . . . . . .
              0 2 2 2 2 1 3 4 3 2 4 3 3 1 1
              9 7 8 2 2 8 2 7 7 4 4 2 2 1 6
              0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
              . . . . . . . . . . . . . . .
              0 1 1 1 1 0 1 2 1 1 2 1 1 0 0
              4 3 4 1 1 9 6 4 9 2 2 6 6 5 8
ITEM RANK =    1  ITEM # =    1  ITEM TOTAL =   97  C =  0.09   C* =  0.04
ITEM RANK =    2  ITEM # =    2  ITEM TOTAL =   94  C =  0.27   C* =  0.13
ITEM RANK =    3  ITEM # =    6  ITEM TOTAL =   85  C =  0.28   C* =  0.14
ITEM RANK =    4  ITEM # =    4  ITEM TOTAL =   81  C =  0.22   C* =  0.11
ITEM RANK =    5  ITEM # =    7  ITEM TOTAL =   76  C =  0.22   C* =  0.11
ITEM RANK =    6  ITEM # =   11  ITEM TOTAL =   68  C =  0.18   C* =  0.09
ITEM RANK =    7  ITEM # =   10  ITEM TOTAL =   65  C =  0.32   C* =  0.16
ITEM RANK =    8  ITEM # =    3  ITEM TOTAL =   65  C =  0.47   C* =  0.24
ITEM RANK =    9  ITEM # =   14  ITEM TOTAL =   65  C =  0.37   C* =  0.19
ITEM RANK =   10  ITEM # =   12  ITEM TOTAL =   58  C =  0.24   C* =  0.12
ITEM RANK =   11  ITEM # =    5  ITEM TOTAL =   43  C =  0.44   C* =  0.22
ITEM RANK =   12  ITEM # =   15  ITEM TOTAL =   40  C =  0.32   C* =  0.16
ITEM RANK =   13  ITEM # =   13  ITEM TOTAL =   39  C =  0.32   C* =  0.16
ITEM RANK =   14  ITEM # =    8  ITEM TOTAL =   32  C =  0.11   C* =  0.05
ITEM RANK     15  ITEM # =    9  ITEM TOTAL =   30  C =  0.16   C* =  0.08

                                P  =   0.509            D*  =   0.369
ITEMS, IN ASCENDING ORDER OF DIFFICULTY

STATEMENTS EXECUTED=   80528

CORE USAGE       OBJECT CODE=   15568 BYTES,ARRAY AREA=  134024 BYTES,TOTAL AREA AVAI

DIAGNOSTICS        NUMBER OF ERRORS=       0, NUMBER OF WARNINGS=       0, NUMBER OF
```

REFERENCES

Airasian, P.W. and Madaus, G.F. A study of the sensitivity of school program effectiveness measures. Report submitted to the Carnegie Corporation, New York. Chestnut Hill, MA: Boston College, School of Education, 1976.

Alker, H.R. A typology of ecological fallacies. In M. Dogan and S. Rokkan (Eds.). Social ecology. Cambridge, MA: MIT Press, 1969.

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

Anastasi, A. Psychological testing. New York: Macmillan Publishing Company, 1976.

Anderson, J.A. Diagnosis by logistic function: further practical problems and results. Applied Statistician, 1974, 23, 397-404.

Anderson, J.A. Multivariate logistic compounds. Biometrika, 1979, 66, 17-26.

Armbruster, B.B., Steven, R.O., and Rosenshine,B. Analyzing content coverage and emphasis: A study of three curricula and two sets (Technical Report

No. 26). Center for the Study of Reading, University of Illinois, Urbana-Champaign, 1977.

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

Baker, E.L. Evaluation of the California Early Childhood Education Program. Vol. 1. Los Angeles, CA: Center for the Study of Evaluation, University of California, 1976.

Barr, R., and Dreeben, R. Instruction in classrooms. In L.S. Shulman (Ed.), Review of research in education, Vol. 5, Itasca, IL: Peacock, 1977.

Berk, R.A. (Ed.) Criterion-referenced measurement: The state of the art. Baltimore, MD: The John Hopkins University Press, 1980.

Berliner, D.C. Studying instruction in the elementary school classroom: Clinical educational psychology and clinical ecomonics. Paper commissioned by the Education, Finance and Productivity Center, Department of Education, University of Chicago, 1978.

Birenbaum, M. Error analysis -- it does make a *difference!* Doctoral dissertation, University of Illinois, Urbana-Champaign, 1980.

Birenbaum, M., and Tatsuoka, K.K. *The use of information from wrong responses in measuring students' achievement* (Technical Report 80-1). Urbana, IL: University of Illinois, Computer-based Research Laboratory, 1980.

Bowles, S. and Levin, H. The determinants of scholastic achievement - An appraisal of some recent evidence. *The Journal of Human Resources*, 1968, *3*, 3-24.

Boyd, L.H., and Iverson, G.R. *Contextual analysis: Concepts and statistical techniques*. Belmont, CA: Wadsworth, 1979.

Brown, B.W., and Saks, D.H. Production technologies and resource allocation within classrooms and schools: Theory and measurement. In R. Dreeben and J.A. Thomas (Eds.), *The analysis of educational productivity*, Vol. 1: *Issues in microanalysis*. Cambridge, MA: Ballinger Press, '980.

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

Burstein, L. Assessing differences between group and individual-level regression coefficients. Sociological Methods and Research, 1978, 7, 5-28.

Burstein, L. The role of levels of analysis in the specification of educational effects. In R. Dreeben and J.A. Thomas (Eds.), The analysis of educational productivity, Vol. 1: Issues in microanalysis. Cambridge, MA: Ballinger, 1980, 119-190. (a)

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, 1980, 158-233. (b)

Burstein, L., Fischer, K., and Miller, M.D. The multilevel effects of background on science achievement at different levels of analysis: A cross-national comparison. Sociology of Education, 1980, 58, 215-255.

Burton, R.B. DEBUGGY: Diagnosis of errors in basic mathematical skills. In D.H. Sleeman and J.S. Brown (Eds.) Intelligent tutoring systems. London: Academic Press, 1981.

Campbell, D.T., and Stanley, J.C. Experimental and quasi-experimental designs for research. Chicago, IL: Rand McNally College Publishing Company, 1963.

Carver, R.P. Two dimensions of tests: Psychometric and edumetric. American Psychologist, 1974, 29, 512-518.

Carver, R.P. The Coleman Report: Using inappropriately designed achievement tests. American Educational Research Journal, 1975, 12, 77-86.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, S., Weinfeld, F.D., and York, R.L. Equality of educational opportunity. (2 Vols.). Office of Education and Welfare. Washington D.C.: U.S. Government Printing Office, 1966.

Comber, L.C., and Keeves, J.P. Science education in nineteen countries. International studies in evaluation, Vol. 1. Stockholm: Almqvist and Wiksell, 1973.

Cooley, W.W., Bond, L., and Mao, B. Analyzing multilevel data. In R.A. Berk (Ed.), Educational evaluation methodology: The state of the art. Baltimore, John Hopkins University Press, 1981.

Cox, D.R. Some procedures associated with the logistic qualitative response curve. In F.N. David (Ed.), Research papers in statistics: Festschrift for J. Neyman. London: Wiley, 1966.

Cronbach, L.J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L.J. Essentials of psychological testing. New York: Harper and Row, 1970.

Cronbach, L.J. (with the assistance of J.E. Deken and N. Webb). Research on classroom and schools: Formulation of questions, design, and analysis. Occasional Paper, Stanford Evaluation Consortium, Stanford, CA, July 1976.

Cronbach, L.J., and Warrington, W.G. Efficiency of multiple-choice tests as a function of spread of item difficulties. Psychometrika, 1952, 17, 127-147.

Cronbach, L.J., and Webb, N. Between-class and within-class effects in a reported aptitude x treatment interaction: Reanalysis of a study by G.L. Anderson. Journal of Educational Psychology, 1975, 67, 717-724.

Day, N.E., and Kerridge, D.F. A general maximum likelihood discriminant. Biometrika, 1967, 54, 313-323.

Dishaw, M. Descriptions of allocated time to content areas for B-C period (Technical Note IV-2b). Far West Laboratory for Educational Research and Development, Beginning Teacher Evaluation Study, 1977.

Divgi, D. R. Calculation of the tetrachoric correlation coefficient. Psychometrika, 1979, 44, 169-172.

Duncan, O. D., Cuzzort, R. P., & Duncan, B. D. Statistical geography: problems in analyzing areal data. Glencoe: Free Press, 1961.

Dyer, H. S. School factors and equal educational opportunity. In Equal Educational Opportunity. Special issue of Harvard Educational Review, 41-59, Cambridge, MA: Harvard University Press, 1969.

Ellett, C.D., Haun, H.C.,Pool, K.W., and Smock, C.D. Planning information study for future Follow Through experiments: Task I response general issues. Mathemagenics Activities Program, University of Georgia, Athens, GA, May 1979.

Filby, N.N., and Dishaw, M. Development and refinement of reading and mathematics tests for grades 2 and 5 (Technical Report III-I). Far West Laboratory for

Educational Research and Development, Beginning Teacher Evaluation Study, 1975.

Filby, N.N., and Dishaw, M. Development of reading and mathematics tests through an analysis of reactivity (Technical Report III-6). Far West Laboratory for Educational Research and Development, Beginning Teacher Evaluation Study, 1976.

Filby, N.N., and Marliave, R. Descriptions of distributions of ALT within and across classes during the A-B period (Technical Note IV-11a). Far West Laboratory for Educational Research and Development, Beginning Teacher Evaluation Study, 1977.

Fisher, C.W., Filby, N.N., and Marliave, R.S. Descriptions of distributions of ALT within and across classes during the B-C period (Technical Note IV-1b). Far West Laboratory for Educaticnal Research and Development, Beginning Teacher Evaluation Study, 1977.

Fisher, C.W., Filby, N.N., Marliave, R.S., Cahen, L.S., Dishaw, M.M., Moore, J.W., and Berliner, D.C. Teaching behaviors, academic learning time and student achievement: Final Report of Phase III-B, Beginning Teacher Evaluation Study (Technical

Report V-1). San Francisco, CA: Far West Laboratory for Educational Research and Development, 1978.

Glaser, R. *The future of testing: A research agenda for cognitive psychology and psychometrics* (Technical Report No. 3). Learning Research and Development Center, University of Pittsburgh, PA, 1981.

Guttman, L. The quantification of a class of attributes: A theory and method of scale construction. In P. Horst, P. Wallin, and L. Guttman (Eds.), *The prediction of personal adjustment*. New York: Social Science Research Council, Committee on Social Adjustment, 1941.

Haney, W. Units of analysis issues in the evaluation of Project Follow Through. Unpublished report, Cambridge, MA: The Huron Institute, 1974.

Haney, W. *A technical history of the national Follow Through evaluation*, Vol. V, *The Follow Through planned variation experiment*. Cambridge, MA: The Huron Institute, 1977.

Haney, W. Units and level of analysis in large-scale evaluation. In K.H. Roberts and L. Burstein (Eds.), *New directions for methodology of social and behavioral sciences*, Vol. 6, *Issues in* aggregation. San Francisco, CA: Jossey-Bass, Inc., 1980.

Hannan, M.T. Aggregation and disaggregation in sociology. Lexington, MA: D.C. Heath, 1971.

Hanson, R.A., and Schutz, R.E. A new understanding of schooling effects derived from programmatic research and development. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, April 1978.

Hanushek, E. The production of education, teacher quality, and efficiency. In Do teachers make a difference?, A report on research on pupil achievement. Office of Education, U.S. Department of Health, Education, and Welfare, Washington, D.C., 1970, 79-99.

Hanushek, E.A., and Kain, J.F. On the value of 'equality of educational opportunity' as a guide to public policy. In F. Mostellar and D.P. Moynihan (Eds.), On equality of educational opportunity. New York: Vintage Books, 1972, 116-145.

Harnisch, D.L., and Linn, R.L. Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational

Measurement, 1981, 18, 133-146.

Harris, C.W., Alkin, M.C., and Popham, W.J. (Eds.). Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles, CA: Center for the Study of Evaluation, 1974.

Henrysson, S. Gathering, analyzing, and using data on test items. In R.L. Thorndike (Ed.), Educational measurement, Washington, D.C.: American Council on Education, 1971.

House, E.R., Glass, G.V., McLean, L.D., and Walker, D.F. No simple answer: Critique of the Follow Through evaluation. Harvard Educational Review, 1978, 48, 128-160.

Howell, D.P., and Rice, C.F. Field coordination Phase III-B (Technical Note II-1). Far West Laboratory for Educational Research and Development, Beginning Teacher Evaluation Study, 1977.

Jenkins, J.R., and Pany, D. Curriculum biases in reading achievement tests (Technical Report No. 16). Center for the Study of Reading, University of Illinois, Urbana-Champaign, 1976.

Jencks, C.S. The Coleman Report and the conventional wisdom. In F. Mostellar & D.P. Moynihan (Eds.), On equality of educational opportunity. New York: Vintage Books, 1972, 69-115.

Jencks, C.S., Smith, M., Acland, H., Bane, M.J., Cohen, D., Gintis, H., Heyns, B., and Michelson, S. Inequality: A reassessment of the effect of family and schooling in America. New York: Basic Books, 1972.

Knapp, T.R. The unit-of-analysis problem in applications of simple correlation analysis to educational research. Journal of Educational Statistics, 1977, 2, 171-186.

Kuder, G.F., and Richardson, M.W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Leinhardt, G., and Seewald, A. Overlap: What's tested, what's taught? Journal of Educational Measurement, 1981, 18, 85-96.

Levin, H.M. A new model of school effectiveness. In Do teachers make a difference? A report on research on pupil achievement. Office of Education, U.S. Department of Health, Education, and Welfare, Washington, D.C., 1970, 55-78.

Lewy, A. Discrimination among individuals vs. discrimination among groups. *Journal of Educational Measurement*, 1973, *10*, 19-24.

Lewy, A. & Chen, M. Longitudinal study of educational achievements, 1971.

Lord, F.M. *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1980.

Lord, F.M., and Novick, M.R. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Madaus, G.F., and Airasian, P.W. The measurement of school outcomes in studies of differential school and program effectiveness. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA, 1980.

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

Madaus, G.F., Kellaghan, T., Rakow, E.A., and King, D.J. The sensitivity of measures of school effectiveness. *Harvard Educational Review*, 1979, *49*, 207-230.

Marliave, R.S., Fisher, C.W., an Dishaw, M.M. *Academic learning time and student achievement in the A-B period* (Technical Note V-1a). Far West Laboratory for Educational Research and Development, Beginning Teacher Evaluation Study, 1977.

Mayeske, G.W., Okada, T., Cohen, W.M., Beaton, A.E., and Wisler, C.E. *A study of the achievement of our nation's students* (DHEW Publication No. (OE) 72-31). Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1973.

Mayeske, G.W., Wisler, C.E., Beaton, A.E., Weinfeld, F.D., Cohen, W.M., Okada, T., Proshek, J.M., and Tabler, K.A. *A study of our nation's schools* (DHEW Publication No. (OE) 72-142), Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1972.

Mehrens, W.A., and Ebel, R.L. (Eds.). *Principles of educational and psychological measurement: A book of selected readings*. Chicago, IL: Rand McNally and Company, 1967.

Murnane, R.J. *The impact of school resources on the learning of inner city children*. Cambridge, MA: Ballinger Publishing Company, 1975.

Muthen, B. Contributions to factor analysis of dichotomous variables. Psychometrika, 1978, 43, 551-560.

Muthen, B. Factor analysis of dichotomous variables: American attitudes toward abortion. In D.J. Jackson and E.F. Borgatta (Eds), Factor analysis and measurement in sociological research: A multidimensional perspective. London: Sage Publications, 1980.

Muthen, B., an Christoffersson, A. Simultaneous factor analyses of dichotomous variables in several groups. Uppsala, Sweden: University of Uppsala, Department of Statistics, 1979.

Popham, W.J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1978.

Porter, A.C., Schmidt, W.H., Floden, R.E., and Freeman, D.J. Practical significance in program evaluation. American Educational Research Journal, 1978, 15, 529-539.

Purves, A.C. Literature education in ten countries. International studies in evaluation, Vol 2. Stockholm: Almqvist and Wiksell, 1973.

Rivlin, A.M. and Timpane, P.M. Planned variation in education: An assessment. In A.M. Rivlin and P.M. Timpane (Eds.), *Planned variation in education: Should we give up or try harder*? Washington, D.C.: The Brookings Institute, 1975, 1-22.

Roberts, K.H., and Burstein, L. (Eds.). *New directions for methodology of social and behavioral science*. Vol. 6. *Issues in aggregation*. San Francisco, CA: Jossey-Bass, Inc., 1980.

Robinson, W.S. Ecological correlations and the behavior of individuals. *American Sociological Review*, 1950, 351-357.

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

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

Singer, J.D., and Goodrich, R.L. Aggregation and the unit of analysis in the National Day Care Study. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, April 1979.

Stanley, J.C. Reliability. In R.L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971, 356-442.

Stebbins, L.B., St. Pierre, R.G., Proper, E.C., Andersen, R.B., and Cerva, T.R. Education as experimentation: A planned variation model. Cambridge, MA: Abt Associates, 1977.

Tatsuoka, K.K., Birenbaum, M., Tatsuoka, M.M., and Baillie, R. A psychometric approach to error analysis on response patterns (Technical Report 80-3). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory, 1980.

Tatsuoka, K.K., and Tatsuoka, M.M. Detection of abberant response patterns and their effect on dimensionalty (Technical Report 80-4). Urbana, IL: University of Illinois, Computer-based Education Laboratory, 1980.

Tatsuoka, M.M. Recent psychometric developments in Japan: Engineers tackle educational measurement problems. Scientific Bulletin, 1979, 4, Dept. of the Navy, Office of Naval Research Tokyo.

Tatsuoka, M.M. Multivariate analysis: Techniques for educational and psychological research. New York: John Wiley and Sons, Inc., 1971.

Thorndike, R.L. Educational measurement. Washington, D.C.: American Council on Education, 1971.

Thorndike, R.L. *Reading comprehension education in fifteen countries*. International studies in evaluation, Vol. 3. Stockholm: Almqvist and Wiksell, 1973.

van der Flier, H. Environmental factors and deviant response patterns. In Y.H. Poortinga (Ed.), *Basic problems in cross-cultural psychology*. Amsterdam: Swets and Seitlinger, B.V., 1977.

VanLehn, K., and Friend, J. *Result from DEBUGGY: An analysis of systematic subtraction errors*. Palo Alto, CA: Xerox Palo Alto Science Center Technical Report, 1980.

Veldman, D.J., and Brophy, J.E. Measuring teacher effects on pupil achievement. *Journal of Educational Psychology*, 1974, *66*, 319-324.

Walker, D.F., and Schaffarzik, J. Comparing curricula. *Reveiw of Educational Research*, 1974, *44*, 83-111.

Wargo, M.J., and Green, D.R. (Eds.). *Achievement testing of disadvantaged and minority students for educational program evaluation*. Monterey, CA: CTB/McGraw-Hill, 1978.

Warm, T.A. A primer of item response theory (Technical Report 941078). Oklahoma City, OK: U.S. Coast Guard Institute, 1978.

Weikart, D.P., and Banet, B.A. Model design problems in Follow Through. In A.M. Rivlin and P.M. Timpane (Eds.), Planned variation in education: Should we give up or try harder? Washington, D.C.: The Brookings Institute, 1975, 61-78.

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

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

Winer, B.J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

APPENDIX B

DRAFT

Deliverable November 1981

State of the Art Methodology for the Design and Analysis of Future Large Scale Evaluations: A Selective Examination

Leigh Burstein

Grant Number

NIE-G-80--0112

CENTER FOR THE STUDY OF EVALUATION
GRADUATE SCHOOL OF EDUCATION
University of California, Los Angeles

Introduction

The following report selectively examines recent developments in quantitative methodology and considers their possible utility in large-scale program evaluations in education. At the outset we limit attention to two specific categories of analytical methods: structural equation modeling and selection modeling and related issues in analysis of quasi-experimental data (non-equivalent control group designs). While these topics, by no means, cover the full range of recent advances in the technology for analyzing quantitative data in large-scale program evaluations, they are representative of the methodological concerns that arise in such investigations, the means analysts propose to deal with the concerns, and the strengths and limitations of primarily technical approaches to resolving ambiguity in evaluation results. As such our examination of these methodological developments is intended to suggest how persons (evaluators, methodologists, agency staff) involved in the design and conduct of large-scale program evaluation might approach decisions about appropriate methodology and its proper use.

Delineation of Relevant Program Evaluations

We further delineate the purview of this investigation by stating the types of evaluation activities and the range of methodological issues to be considered. We are concerned with field-based investigations of large-scale programs typically approved by legislative actions and implemented (or to be implemented) by governmental agencies. Both evaluations of on-going programs (e.g., Title I) and of various forms of social experiments (e.g., Negative Income Tax experiments) are relevant to the present discussion (Cook (1981) restricts his attention to the former). The domain also encompasses both well-defined programs (i.e., those with a discrete number of specific program alternatives such as the various models in operation in Planned Variation Follow Through) and broad-based educational reforms as represented by Title I, the Emergency School Aid Act (ESAA), and bilingual education. (A related paper (Burstein, 1981) focussed strictly on evaluations of well-defined programs).

Types of Evaluation Questions

The limits placed on the evaluation activities of interest are in the kinds of questions one seeks to answer and the form of data collection in the evaluation. Cook (1981) discusses six types of questions that evaluators try to answer:

1) Who are the clientele and service providers and to what extent are target groups among the clients? (Demography)
2) What are the delivered services and the contexts in which services are received? (Implementation)
3) How do program services affect clients in both expected and unexpected ways? (Effectiveness)
4) How are other elements (teachers, schools, families, etc.) of the educational system affected by the program services? (Impact)
5) Why do program services affect outcomes in the way they do? (Causation)
6) What are the costs of the services and how cost-effective are different ways of achieving a particular result? (Economic costs)

The questions about effectiveness, impact, and causation are central to our examination. To be comprehensive, investigations of these types of questions require information about the characteristics of the program, its clients and participants and the context in which it is implemented, the educational and social processes (intended and actual) occurring within program sites, and the outcomes of programs at various levels (student, teacher, classroom, school, community, etc.) of the educational system. Conceptual and analytical machinery are then employed to elucidate the linkages and connections among the various sources of information.

Types of Data Collections

In the past, most large-scale field evaluations of educational programs collected mainly "quantitative" measures of program characteristics and outcomes largely derived from survey questionnaires completed by clients and other relevant program participants (e.g., teachers, principals, parents), limited interviews with program personnel and observations of program activities (e.g., Stallings and Kaskowitz, 1974), and paper-and-pencil measures of cognitive and affective outcomes. Data were collected from multiple sites for each variant of the program to achieve a given degree of information about program variation and a sufficient number of observations for statistically powerful tests of program effects.

Recently, however, data collection in even large-scale program evaluations has taken on an increasingly "qualitative" character. Extended case studies were conducted in either a subset or all sites in a number of recent large-scale evaluations (e.g., Title I Parent Involvement Study conducted by SDC; Study of the Longitudinal Effects of the California Early Childhood Education Program conducted by CSE, the Rand Study of Federal Programs Supporting Ed cational Changes, the evaluation of Curriculum Development Projects in Science Education conducted by CIRCE). At the least, the inclusion of case studies in these evaluations provide a richer picture of program process than was obtainable from strictly questionnaire information. And, as methods for synthesizing multiple case studies and integrating qualitative and quantitative information improve, qualitative methods will play an increasingly more prominent role in the reportoire of evaluation activities previously concentrated on less dense forms of data collection.

Despite the increasing role of qualitative methods and our positive attitude about their central role in future evaluations, the remainder of the paper will restrict attention to developments in quantitative methods from multi-site investigations using questionnaire, interview, test and perhaps small-scale observational data. We impose this restriction for two reasons. First, the analytical developments considered are appropriate primarily for the more traditional kinds of quantitatively oriented studies. Second, others (e.g., Daillak & Alkin, 1981) are more capable at this point of stating the case for qualitative methods.

Overview of the Report

The remainder of the report will proceed as follows. First, a general overview of current perspectives on the design and conduct of large-scale program evaluations is presented. The intent is to explain why the climate for future large-scale evaluations is conducive to the introduction of improved methods of analysis. Second, two specific categories of analytical methods (structural equation modeling, and selection modeling/analysis of non-equivalent control group designs) are considered. The basic conceptual and analytical foundations for each method are described, issues that motivate its use in program evaluations are delineated, and specific strengths and weaknesses of each method in program evaluation contexts are identified.

Current Perspectives on Design and Analysis in Large-scale Program Evaluations

There are strong signs that large-scale educational evaluation has witnessed the end of an era. From the late '60's and throughout the 1970's, the federal government, under legislative mandate, mounted major evaluations of just about every conceivable educational program. Wargo (1977) points to 110 major evaluations of federal educational programs funded by the Office of Planning, Budgeting, and Evaluation of the Office of Education at a cost of over $80 million during the 1971-1979 period. The figure does not even include all the major evaluations done by the Office of Education, much less NIE and other branches of HEW.

Many of these large-scale multiyear studies have been highly visible in the educational community though their direct influence on legislative action is less clear (Barnes & Ginsberg, 1979; Cohen & Garet, 1975; Cross, 1979 Wisler & Anderson, 1979). In most cases, the debates about the quality and merits of these evaluations have been heated. This has especially been the case for evaluations of compensatory programs such as Head Start (e.g., Cicirelli et al., 1969, 1971; Smith & Bissell, 1971), Project Follow Through (Anderson, 1976; Cline et al., 1974; Haney, 1977a, 1977b; House, Glass, McLean, & Walker, 1978; Stebbins et al., 1977), and Bilingual Education (AIR, 1979; Center for Applied Linguistics, 1979). The literature on evaluations of these programs is replete with critiques, reanalyses, and secondary analyses, not to mention the often self-serving attacks from program advocates and critics.

Signs of Change.

Emphasis. There are clear signs, however, that the large-scale evaluations of the 1980's may well be different. First, recent scholarly (e.g., Cook, 1981; Cronbach, 1978; Cronbach & Associates, 1980; House, 1977, 1979; Raizen & Rossi, 1981) and policy (e.g., Boruch & Cordray, 1980) contributions provide well-reasoned accounts of the complexity of program evaluations in highly politicized contexts and persuasive arguments for different views of evaluation's role in the formation of social policy. These writings urge that less emphasis be placed on the traditional social science/experimental design paradigm for impact evaluation while more effort be devoted to describing and explaining the processes of educational programs and their consequences over a broad range of outcomes. The overly simplistic overall program impact question (i.e., does program A affect pupil outcomes?) that guided so many of the OPBE funded studies (e.g., ESAA (Coulson et al., 1977); Follow Through (Stebbins et al., 1977); and Bilingual Education (AIR, 1979)) appears to be on the decline.

Instead, recent evaluations involve more direct efforts to investigate and describe the consequences (intended and otherwise) of educational programs. This "information" as characterized by Cronbach et. al. (1980) involves a "move away from stand-alone evaluations of programs and toward a more synoptic view ~~of~~ the numerous programs that address the same social programs" (p. 72-73) and they urge that evaluations employ multiple studies using different strategies to investigate subquestions and that the evaluation plan evolve as individual studies expose uncertainties more clearly. The NIE Compensatory Education Study (NIE, 1977) and the evaluations of services to handicapped children under Public Law 94-142 (Bureau of

Education for the Handicapped, 1978) are clear examples of this type of evaluation.

This shift in evaluation emphasis is a logical response[1] to the findings that variation in implementation within a program is generally greater than between programs (Stebbins et al., 1977), new program "treatments" are quickly diffused to non-participating groups (schools, etc.) (Coulson, 1978) and that the effects that are discerned depend on the characteristics of the program processes "as implemented" rather than on the ascribed program characteristics (Cook, 1981; Cronbach, 1978; Cronbach & Associates, 1980; Rogosa, 1978). Under such conditions, only those evaluation activities that delve beneath the surface descriptions of programs can be expected to generate quality information for policy formation.

Methodological improvements. Clearly, the impetus for change in the conduct of large-scale educational evaluation exists. The philosophical, theoretical, and political bases for the changes have been and are being articulated. Under such conditions, the climate for evaluation in the 1980's is quite open to new designs and strategies for evaluating the effects of educational programs. The task of defining these designs and strategies and illustrating their worth remains.

Fortunately, it is unnecessary to begin from scratch in the design of large-scale evaluations for the 1980's. While actual educational evaluations over the past decade, for the most part, utilized pre-1970's technology (quantitative methodology, psychometric methods), investments of resources in basic research on methodology and measurement during the 1970's led to substantial improvements in the state of the art.

The relatively unsophisticated applications of experimental, quasi-experimental, and non-experimental methods that led to the findings of the Coleman report (Coleman et al., 1966) and of early Head Start and Follow Through evaluations need not be repeated. Better, more sensitive quantitative methodology is now available and is more suited to the shift in emphasis in large-scale evaluations.

The same can be said for the measurement of program outcomes and processes. Approaches for developing program sensitive test instruments as well as a broader view of the range of program outcomes are currently on the evaluation agenda. The investment of resources to obtain more intensive and descriptive measures of program implementation and processes appears to be a standard feature of recent large-scale evaluations (e.g., the Title I Parent Involvement Evaluation conducted by Systems Development Corporation). These measurement strategies should facilitate more useful evaluations. Better methods of knowledge and data synthesis (e.g., recent work by Glass and Light) should also contribute to better evaluations.

Basis for Methodological Improvements

The special issue of the Journal of Educational Statistics on the Emergency School Assistance Act (ESAA) Evaluation (JES, 1978; see especially Rogosa, 1978) and Cronbach's report on designing educational evaluation (1978) provide documentation of key evaluation methodology issues and help to motivate our general concerns. The basis for our investigation into evaluation methodology is in part the following set of general premises:

(1) Evaluation is inevitably an empirical enterprise, "examining events in sites where the program is tried and the reactions and subsequent performance of the persons served .... (as such it) is typically identified with the application of social science methods: observation, measurement, and/or use of informants." (Cronbach, 1978, pp. 25-26).

(2) "The success of an evaluation effort should be measured by its social usefulness or utility .... Technical decisions should not be made independently of the political and social context of an evaluation. The central question is: How can we design, analyze and report evaluations so as to make them maximally useful?" (Rogosa, 1978, p.80, emphasis added).

(3) "Evaluators are unwise to collect data only on pretest and posttest achievement measures or conduct analyses that only determine the statistical significance of the overall treatment effect. Additional data on process, and on program realization, are essential for adequate descriptions of programs operating in complex settings." (Rogosa, 1980, p. 81).

(4) The analytical strategies in program evaluations should be adapted to the substantive problems under investigation rather than adapting the evaluation of program impact to fit the analytical methods. Natural designs and analysis should evolve from the structure and function of the program. (Burstein, 1980).

(5) Program evaluation is typically carried out within a multilevel educational context. Program activities occur in the groups (classrooms, schools, etc.) to which an individual belongs. These groups

influence the thoughts, behaviors, and feelings of their members. (Burstein, 1980).

(6) Educational interventions are typically implemented within ongoing programs. They vary in "fit" with existing activities and predilections and vary in duration. Interventions in social settings are inherently dynamic activities.

There are more specific methodological corollaries to these general premises:

(1) "No one level is uniquely responsible for the delivery of and response to educational programs ... confining substantive questions to any one level of analysis is unlikely to be a productive research strategy" (Rogosa, 1978, p. 83). Thus, attempts to answer questions about the effects of educational programs require analyses at and within the levels of the educational hierarchy (Burstein, 1980).

(2) Even when one starts with a controlled experiment with random assignment, features of the experimental design break down through processes of attrition, contamination, and differential penetration of the treatment. Under such conditions, quasi-experimental forms of adjustment and control are inevitably necessary and thus should be anticipated as part of the evaluation design.

(3) In the course of an educational program, students are members of multiple groups (e.g., classes). The features of these group contexts and the consistency of student's educational experiences within them over time warrant consideration for dynamic modeling of program experiences (Burstein, 1981; Tuma, Hannan, & Groenfeld, 1978; Rogosa, 1980).

(4) In field experiments with well-defined treatments, the variation in the fidelity of program practices with teacher (school, etc.) predilections and skills leads to a continucus range of program processes. Under these conditions, modeling the intervention as a dichotomous rather than a continuous event is an insufficient approach for investigating program effects (Burstein, 1981; Cronbach, 1978; Rogosa, 1978).

(5) Even when random assignment occurs at some aggregate level (e.g., school), the variation in the treatment effects for students within aggregates needs to be investigated, especially in terms of its consequences for the equalization of educational opportunity.

(6) Programs have multiple effects. Multiple measurement is needed to encompass intended and unintended effects (desirable or undesirable), (Cronbach, 1978, p. 26).

Fortunately, one can point to specific bodies of methodological work that are responsive to both the general perspectives and the accompanying methodological corollaries. In the following sections we will elaborate the connections for a selected set of methodological strategies.

Examination of Specific Analytical Developments

The analytical methods to be examined represent broad areas of methodological concerns that first developed within social science research in general. To understand why this is both an obvious and proper starting point, one need only consider the criteria used to delineate our relevant universe of large-scale program evaluation. In particular we are interested in design and analytical problems in evaluations that fit the following description:

(1) The evaluation should have been conducted on a distinct funded educational pr)gram(s) rather than be a general shift in the behaviors of an educational system. There must have been some form of intervention, innovation, or change in the ongoing educational program.

(2) The evaluation must have involved multiple sites of each presumably distinct program type.

(3) The program must have been implemented (i.e., the main program activities must operate) at the level of the school or lower.

(4) Both outcome and program process data must have been collected during the course of the evaluation.

(5) Outcome data must be available over multiple time points.

(6) Good documentation of the original evaluation must exist.

The above delimiters eliminate evaluations which are short-term efforts, have a limited number of sites, or are of programs presumably constant over all schools in a district. These criteria include evaluations of well-defined program interventions such as in provided by a specific Head Start or Follow Through model, interventions that are less specific in program prescription but nonetheless are assigned to "sites" in a systematic manner such as by random a signment (e.g., the ESAA Evaluation) and more pervasive social interventions where participants are essentially all persons with a prescribed set of characteristics (e.g., Title I, Bilingual Education).

To gain a better perspective on the kind of study situation evnisioned consider the following modified version of the conceptual framework for investigating the impact of educational reforms outlined in Burstein (1981). One starts by identifying the specific elements of educational and social systems in which programs are introduced and the processes and outcomes that result. The elements are the characteristics and attributes of individual students, families, groups of students, teachers, classes, groups of teachers, schools, and communities. The processes are developmental, instructional, curricular, psychological, interpersonal, and social. Both elements and processes can take on either static or dynamic properties though the latter are more likely in school settings, especially those with large numbers of poor children participating in school reform programs.

A general model containing the essential elements and processes of the conceptual framework is as follows. The interrelations among five distinct classes of variables are incorporated in the model: program instruction, schooling context (class, school, community, etc.), stedent entering characteristics, and student performance. Each class may represent many distinct variables (or sets of variables). For example, "instruction" refers to the various characteristics of the instruction a student receives in a specific classroom or school. Particular teacher attributes (e.g., warmth, enthusiasm, clarity of presentation) and instructional processes (e.g., structure; grouping, pacing, types of reinforcements, teachers' questioning behavior, quality and variety of instructional materials) both fit under the instruction rubric. Certain aspects of the instructional practices also provide evidence about the degree of program implementation. Nonetheless, any measure of program implementation would still fall within the "instruction" category for present purposes.

The term student "performance" is meant in the broad sense; the full range of educational, social, and psychological outcomes fit under this general rubric. The restriction to student outcomes could be broadened to include other units (teachers, classes, schools), but not without making the task of generating the framework even more unwieldy than it will appear here.

The role of schooling context in the model is multifaceted. Its most proximal manifestations are in the classroom where the program is implemented. For example, the overall level and heterogeneity of ability in a class places constraints on instructional content, organization, and management. The consequences of these constraints vary for different reform programs. Class heterogeneity places a strain on time and resources in individually prescribed educational programs. Decisions about the pacing of instruction become more difficult in programs emphasizing large group instruction.

The student's role within the classroom is also directly influenced by its composition (Burstein, 1980b; Firebaugh, 1980; Webb, 1980). There is obviously a complicated balance between having classmates compatible in ability and temperament versus having peers that are more or less able and/or have contrasting personalities. Either combination might foster intellectual, social, and psychological growth under the "right" conditions. Here, again, programs with different emphases and organization might interact differentially with class composition, making a given student's role more comfortable or stressful.

There are also other elements of context provided by the class, school and community environment for the program. Sirotnik and Oakes (1981) provide a particularly comprehensive discussion of the possible components of schooling context.

The pattern of relationships depicted in Figure 1 include the following:

(1) Students are eligible for the program and are selected on the basis of entering characteristics.

(2) Student entering characteristics (ability, "preferred learning style", motivation to learn, "preparation for learning") affect performance at any point in time.

(3) Entering characteristics interact with program characteristics to give certain students relative advantages in certain programs (e.g., low ability students benefit from relatively higher levels of teacher control and direction for language and mathematics mechanics).

(4) Programs interact with school personnel characteristics (preferred style, personality, authority relationships, cohesiveness).

(5) Schooling context (ability distribution, personality, presence/absence of demanding/disruptive students, orderliness at class or school level) affects instruction (emphasis, amount of material covered, organization, program delivery).

(6) Students' shared educational and social experiences in classrooms and schools depend on student entering characteristics, instruction, schooling context and program characteristics.

(7) Students from same class in year 1 may be assigned to different classes in year 2 or may leave the school.

(8) Students not present in year 1 may enter school (and thus program classes) during year 2.

(9) Implementation of programs may differ for year 2 from year 1.

(10) Instructional (program) characteristics e.g., teacher "style", organization) may differ from year 1 to year 2 and effect of instruction (program) year 1 followed by instruction (program) year 2 is not necessarily additive.

(11) Contextual characteristics may differ from year 1 to year 2.

(12) Conditions (1) - (5) hold for year 2 in similar fashion as for year 1.

(13) Program differs from "normal" standard instruction and may interact. Though instruction of Type A may be better than instruction of Type B, instruction of Type B might be better for students following participation in the program than Type A would be.

The two areas of analytical developments to be discussed below become relevant in a program of the type described above for several reasons. First, eligibility for program participation typically depends on specific ascribed characteristics (e.g., poverty, bilingualism, ethnicity). Even in nominally "experimental" investigations, selection for participation may have non-random aspects at some level as in the case where the program is randomly assigned to a sample of schools from a pool of volunteers. A further complication is the non-stable participant sample; students enter and leave classrooms, teachers and schools drop out of programs for various reasons.

A second feature requiring analytical attention is the sheer number of elements that potentially enter a comprehensive picture of program processes and outcomes, the complexity of their interrelation, and the inherent problems in measuring key variables by the kinds of questionnaire, interview, observation and test data typically used. All of the elements of model specification from a clear understanding of the question of interest through identification and operationalization to appropriate analyses and interpretation have a bearing on the fidelity of the evaluation conclusions to the program's actual consequences.

To a certain degree, these features align with the two analytical developments to be considered below.

Non-Equivalent Control Group Designs/Selection Modeling

From the inception of the large-scale educational evaluation efforts of the 1960's, evaluators have tried to employ the paradigm for experimentation in the field investigations. With rare exception, however (see Boruch, 1974), investigators quickly found themselves in the midst of non-experimental or at best quasi-experimental studies wherein all the best intentions about random assignment went unfulfilled.

From a methodological perspectice, consciousness about the inadequacy of analytical methods in these investigations can be traced back to Campbell and Erlebacher's (1970) lament (perhaps complaint is the better term) that regression artifacts in quasi-experimental evaluations were causing compensatory education to look harmful. While certain aspects of the original Campbell-Erlebacher critique have been found to be less generally applicable than originally believed, the design constraints that bothered them remain at the center of current analytical concerns.

Basic analytical issues. Reichardt's (1979) and Barnow, Cain and Goldberger's (1980) discussions of the problems in analyzing non-equivalent control group designs are a particularly helpful starting point for our examination. As Reichardt points out, the main issue is the effect of uncontrolled selection on the estimation of program effects. When subjects are randomly assigned to programs (or non-program), groups can be considered initially equivalent though the equivalence can be vitiated if there is differential attrition. Without random assignment program groups would not be expected to equal even in the absence of a program effect. Thus, in order to "equate" non-equivalent groups, it is necessary to adjust or control for initial differences.

The analyst this juncture invariably recognizes that the task at hand is to (a) identify the selection process underlying group membership (program, non-program) and (b) include the variables that determine selection in the analysis of program effects. Ideally, this analytical strategy would control for the effects of initial differences.

Until recently the statistical method typically employed by analysts in quasi-experiments was the analysis of covariance (ANCOVA), which is essentially a linear regression of program outcomes, Y on program status Z, (e.g., 1 = in program, 0 not in program) and pre-program true ability W^2. Thus the "ideal" analytical model is represented by (1) below:

$$Y = \alpha Z + W + \varepsilon \quad , \qquad (1)$$

where α is the estimate of program effect, Z, and W is the covariance adjustment for true initial differences.

But as is well-known, W is unobservable. Under these conditions Barnow, Cain and Goldberger (1980) ask "How may the evaluator persuade an interested audience that the measured effect of Z on Y is free of any contamination from a correlation between Z and W, given that W is not available as an explanatory variable?" (p. 47). Their answer to their own question is that "unbiasedness is attainable when the variables that determine treatment assignment are known, quantified and included in the equation." (Barnow, et. al., 1980, p. 47. See also Barnow, 1975; Cain, 1975; and Goldberger, 1972). Thus if one has an observed variable t that was used to determine group assignment (in general t will be a score based on a composite of variables, some of which may be correlates of W), then t may be used to replace W as the explanatory variable in (1):

$$Y = V_1 Z + \beta_2 t + \varepsilon^* \qquad (2)$$

Under conditions to be specified, β_1, in equation (2) would be an unbiased estimate of the program effect α. Thus either W or t will remove the contamination which leads to "selectivity bias".

But the question arises about whether the selection process can be known precisely (i.e., one is unable to quantify t). In this case, investigators have settled for a set of variables, X, that serve as proxies for W. The X's may also include variables which enter t. The equation to be estimated is then

$$Y = \gamma_1 Z + \gamma_2 X + \varepsilon^{**} \quad . \qquad (3)$$

Equation (3) is essentially the standard ANCOVA model as employed in the analysis of quasi-experimental data. Unfortunately, an estimate of γ_1 will in general be a biased estimate of the true program effect α. Statistically, this bias depends on the covariance of Z and W conditional on X. Moreover, contrary to Campbell and Erlebacher's (1970) assertion, the bias may be either positive or negative. Investigations by Goldberger (1972), Barnow (1973), Cain (1975), Cronbach, Rogosa, Floden, and Price (1977) and Bryk and Weisberg (1977) clearly demonstrate this property.

To better understand the ramifications of the inability to observe true preprogram ability (W) and/or to accurately quantify the selection process (t), we consider the sources of biases in estimation of program effects when the ANCOVA model is employed with nonequivalent groups. Reichardt (1980) discusses seven sources, most of which are pertinent to this inquiry.

The problems due to errors in measuring the covariates (the X's in equation (3)) are the most frequently examined source of bias. Even when measurement errors are random, they lead to attenuated estimates of covariate effects and thus result in an underadjustment for pre-existing differences between different programs. The errors in the covariate cause the treatment effect estimate from ANCOVA to converge toward estimates from an ANOVA which completely ignore pre-existing group differences.

The second source of bias in ANCOVA is the possibility of differential growth rates among identifiable subpopulations under conditions where subpopulation membership is related to program assignment. Though individuals from different subpopulations may be the same initially, their later differences may be attributed to differences in maturation. In this case, growth invalidates ANCOVA because within-group growth does not completely account for between-group differences in growth.

According to Reichardt, related sources of bias due to changes between the time of program entry and measurement of program outcomes which are irrelevant to the treatment are trait instability and the changing structure of behavior. Trait instability refers to differential variability (fluctuation) in scores over time as opposed to average mean differences. The changing structure of behavior refers to the possibility that the processes that account for given naturally occurring behaviors vary over time with different characteristics and processes becoming disproportionately important at various times. (Cronbach et al (1977) discuss this source in some detail.)

Other complications identified by Reichardt include (a) operationally unique pretests and posttest (i.e., even though the measure of initial status and final performance is nominally the same, they are operationally distinct as different abilities and skills are tapped at different points in time); (b) non-linear regression lines (not properly incorporated in the model) and non-parallel regression lines (due to treatment interaction effects, floor and ceiling effects, differential growth between groups, or between group differences in the reliability of the covariates).

Reichardt (1980) describes four approaches for ruling out selection differences as a rival explanation for program effects. The first three (namely, developing a causal model of the posttest, developing a causal model of the assignment process, the Cronbach et. al. (1977) combination of the two approaches) are basically elaborations on the identification of W, t, or both as described earlier. One essentially adopts a broader, theoretically grounded and empirically estimated model of how posttest behavior is expected to vary in the absence of the program (modeling the posttest; Cronbach et. al. call this identifying the "ideal covariate"), how individuals are assigned to "treatment" groups (modeling the assignment process; or identifying the "complete discriminant" in Cronbach et al.'s terminology) or do both. After determining a specific approach, there are still questions about appropriate analytical machinery to adjust for measurement errors and estimate W and t appropriately. The sheer complexity of the adjustment has led some investigators to recommend the use of procedures derived from the work of Joreskog (1970, 1973, 1974, 1977, Joreskog and Sorbom, 1976, 1978) for the analysis of covariance structures. These methods attempt to simultaneously correct for the effects of measurement error and irrelevance in multiple covariates. We withhold further discussion of these techniques to the next major section of our report.

Value-added analysis. The fourth approach discussed by Reichardt (1980) is the modeling of change or growth. Promising work on this topic has been carried out by Bryk and Weisberg (Bryk, 1977; Bryk and Weisberg, 1976; Bryk, Strenio , and Weisberg, 1980; Strenio, 1977; Weisberg 1978). They introduced a variety of analytical methods for estimating the "value-added" by program participation. Their value-added analysis is built upon the notion that educational programs are dynamic interventions in natural growth processes. Thus Bryk and Weisberg first modeled natural growth processes and then assessed program impact on the processes.

The basic idea underlying Bryk-Weisberg value-added procedure is to compare average observed growth between pre- and post-test with an estiamte of the amount expected in the absence of an intervention.

To employ their techniques, one needs to have pretest (Y_{1i}) and post-test data (Y_{2i}) on a sample of individuals as well as the time (calendar dates t_1 and t_2) at which observations were obtained and the age (a_{i1}, a_{i2}) of each individual at these times. In the more general case, one would also obtain information on other background variables (X_i). Their methods also seem to be applicable whether treatment is represented by a discrete group membership variable (treatment A vs. treatment B) or by a set of variables describing program and instructional differences (e.g., explicit charicteristics of instruction, schooling, context, and program implementation).

Bryk and Weisbergs's general model can then be expressed as

$$Y_i(t) = G_i(t) + R_i(t) \qquad (4)$$

$$G_i(t) = \pi_i a_i(t) + \delta_i \qquad (5)$$

$$\pi_i = \theta_0 + \sum_{j=1}^{J} \theta_j X_{ij} + \varepsilon_i \qquad (6)$$

In (4) above, $G_i(t)$ and $R_i(t)$ represent systematic growth and random components respectively. π_i and δ_i are slopes and intercepts of individual growth curves, $a_i(t)$ is the age of individual i at time t. The X_{ij} are the values of the jth background variable for subject i, $\underset{\sim}{\theta}_j$ are the corresponding coefficients and ε_i are unmeasured determinants of individual growth rates. Given one of several choices of assumptions about error structure (e.g., $E[R_i(t)] = 0$; Var $(R_i(t)) = \sigma_r^2$, constant over all subjects and times; R_i independent of t, π_i, δ_i, and any R_i; $E(\varepsilon_i|X_i) = 0$; Var $(\varepsilon_i|\underset{\sim}{X}_i) = \sigma_\varepsilon^2$ and Cov $(\varepsilon_i, \underset{\sim}{X}_i) = \underset{\sim}{0}$), one then estimates the value-added by first regressing pretest on age and its interactions with background variables to determine estimates of individual growth rates (π_i^*) and then calculates a value-added for an individual using the expression

$$v_i = Y_i(t_2) - Y_i(t_1) - \pi_i^*\Delta_i, \qquad (7)$$

where Δ_i represents the time interval between pretest and posttest. The average of the individual value added,

$$V = \frac{\sum_{i=1}^{n} v_i}{n} \qquad (8)$$

is then an estimate of program impact.

Byrk and Weisberg's procedures appear seductively simple and broadly applicable. One models the growth process as best one can from relevant background variables and the time span over which the program measurements are obtained then attributes the remaining average increment in performance to the program. In their most recent article (Bryk et al., 1980), extensions of the basic value-added analysis model to cases where errors in regression models are heteroscedastic, growth is non-linear, comparison group data are available, when programs are administered to non-randomly formed groups of individuals, and when aptitude-treatment interactions are believed to exist are discussed.

Important limitations of the value-added procedure are also indicated by Bryk _et al_. (1980). The problem of a shifting metric for measuring growth over time cannot be alleviated through value-added procedures. Whether it is simply a matter of the restandardization of scores at different age and grade levels or the more serious (analytically, at least) concern that the component skills accentuated at different ages vary, the basic complication falls outside the purview of a modeling procedure of this type.

Another limitation is the inability of the lone value-added model to deal with the lack of monotonicity of growth that occurs in schooling data with multiple years of schooling separated by summer vacations. In our companion report (Miller, 1981), a rudimentary example of this non-monotonicity arises in the Beginning Teacher Evaluation Study (BTES) data. Maddahian (1981) showed that this occurred for other BTES measures and others (e.g., Klibanoff & Haggart, 1980) have uncovered similar examples in other evaluation studies. It is not inherently impossible to apply the value-added approach to more complex growth models; it is just unclear at present how one converges substantively on an adequate model for these more complex dynamic processes.

There is no mention in the Bryk-Weisberg work of how the investigator is to alleviate the problem of measurement errors in explanatory variables.

While the concentration on a single group model (no comparison group) seemingly removes the concerns about differential attenuation of estimates the t.o-stage estimation process (estimate growth from pretest and predict growth increments to subtract from posttest) would appear to place greater demands for precise estimation not likely to be met by the current value-added approach. In principle the model should work best during periods when individuals are experiencing substantial observed growth which suggests that the technique is most suitable for the study of programs for younger children. But outcome measures are notoriously less reliable and stable during the preschool years and early grades of formal schooling than in later years.

Similarly, from a modern perspective, it is advantageous to be able to model program processes and examine their effects directly rather than rely simply on program participation as the indicator of program effects. As Bryk et al. (1980) demonstrate, the value-added approach can be used to estimate the effects of program characteristics on program outcomes (i.e., the value-added for a given site). Yet here, too, the errors in measuring program process characteristics as opposed to, say, ascribed individual and program characteristics are likely to inadequately reflect the true state of affairs.

Finally, there is no provision in the current literature on the value-added approach to deal with multiple measures of growth. Presumably, analysts must choose some means of arriving at a single growth measure (e.g. some form of composite) before proceeding with the value-added analysis. The alternative is to generate a series of value-added estimates, one for each combination of pre- and posttests. Our sense is that the former will typically be less than satisfactory because of the changing character of the ideal composite over time. The latter quickly becomes unwieldy unless a reasonable scheme of interpreting the pattern of effects can be determined (e.g., see Weisberg, 1978).

In conclusion we judge the value-added approach to be a useful addition to the complement of analytical strategies for evaluating program consequences. However, the biases associated with measurement errors, changing metrics and the changing structure of behavior linger and may, in certain respects, be exacerbated. Nor is the multiple measures of outcome programs adequately considered. Nonetheless, if investigators do choose to employ the multiple analysis strategies perspective advocated here, the value-added approach will be a wise choice for inclusion in a broad range of evaluation situations.

Selection modeling. Another recently developed set of analytical approaches for dealing with selection bias can be traced to evaluations of social experiments on welfare reform (Rossi & Lyall, 1976; Stromsdorfer & Farkas, 1980). Economists working on these evaluations developed methods for adjusting for selection effects in estimating the effects of interventions. Volume 5 of the Evaluation Studies Review Annual (Stromsdorfer & Farkas, 1980) is the most comprehensive published source on selection modeling methods. Representative papers from several of the major contributors (e.g., Hausman, Heckman, Goldberger) are included along with useful discussions of the issues by the editors (Stromsdorfer & Farkas, 1980), and by Barnow, Cain, and Goldberger (1980). However, this work is rapidly developing and even recent synthetic reviews by Muthen (Muthen, 1981; Muthen & Joreskog, 1981) cannot keep up with the latest technical nuances. In addition a whole set of seemingly related techniques developed by sociologists (e.g., Tuma & Hannan, 1978; Tuma, Hannan, & Groenveld, 1978) for dynamic modeling with panel data are not even considered by the economists.

We will not attempt to describe all the particular analytical developments in our discussion of selection modeling. Instead, we try to indicate the ways in which the methods are designed to alleviate specific problems in the analysis of quasi-experimental data, point out the broad categories of analytical approaches that are currently available, and attempt to pinpoint the set of problems left unresolved by these methods. And, although we find the methods of Tuma and Hannan potentially valuable for longitudinal evaluations of social programs, the discussion will concentrate on the econometric work.[3]

The general problem that motivates the selection modeling work is the selectivity bias that results when individuals (or, for that matter, aggregates of individuals such as schools) are self-selected (non-randomly selected) into experimental and control groups (or into different program types) or when data on the study sample are non-randomly missing (see our earlier discussion of work by psychologists on this topic (i.e., work reviewed by Reichardt , 1979). According to Stromsdorfer and Farkas (1980), "the realization that the difficulties associated with self-selection, censored samples (where some variables are unmeasured for certain individuals in the sample), truncated samples (where all variables are unmeasured for certain individuals who should be in the sample), and limited dependent variables (variables restricted to some subset of values: for example, weeks worked, which must be zero or above or the probability of being employed, which must lie between zero and one) all have a common foundation" (p. 14) was perhaps the most important statistical development in social science methodology during the 1970's. This realization led investigators to develop methods for incorporating analytical procedures for handling self-selection, censored and truncated samples, and for limited dependent variables within the general analytical model for estimating program effects.

The general analytical procedures involved in econometric selection-modeling can be sketched as follows. (This discussion draws heavily from Barnow, Cain, and Goldberger (1980), Goldberger (1979), and Muthen and Joreskog (1981).) Because of non-random assignment to program it is necessary to incorporate information about the selection process into the equation for estimating program effects. Thus, equation (3) for program outcomes,

$$Y = \gamma_1 Z \text{ r } \gamma_2 X + \varepsilon^{**} \qquad (3)$$

(remember Z represents program; Z=1 for program participated and Z=0 for group comparison) needs to be supplemented by an equation for selection into the program. A selection equation with Z as the dependent variable is specified and restrictions are placed on it to remove pre-existing differences between program and comparison gorups from the estimates of the treatment effect (γ_1 in (3)). The restrictions on the selection equation appear to be of two types. First, there must be variables that determine selection that do not affect outcome. Thus, there must be variables necessary to account for Z that are not among the X's from equation (3). Second, the functional form of the relation between X and W (true ability as identified in equation (1)) and a non-linear relation between Z and X are specified. This leads to a non-linear functional form of X in the outcome equation that is necessary to control for any relationship between Z and W that is not controlled by X.

In more formal terms we begin with three observable variables (Y, X, Z), two unobservable variables (W and t, the true selection variable; these two are anaologous in many respects to Cronbach et al.'s ideal covariate and complete discriminant) and various disturbances for the equations. Then

$$Z = \begin{cases} 1, \text{ if } t > 0 \\ 0, \text{ if } t \leq 0 \end{cases} \qquad (9)$$

and, as stated earlier program outcomes are determined by

$$Y = W + \alpha Z + \varepsilon_0 \qquad (1)$$

where ε_0 (ε in original version of equation (1)) is normally distributed, independent of W and Z, and has expectation zero and standard deviation σ_0. the relations among X, W, and t prior to selection and program participation are given by

$$W = \theta_1' X + \varepsilon_1 \tag{10}$$

$$t = \theta_2' X + \varepsilon_2 \quad , \tag{11}$$

where θ_1' and θ_2' are coefficients relating X to W and t, and disturbances ε_1 and ε_2 are bivariate-normal, uncorrelated with X and ε, have standard deviations σ_1 and σ_2 and covariance σ_{12}. Thus, W and t may be related via X or through correlated disturbances. Substituting from (10) into (1) yields

$$Y = \theta_1' X + \alpha Z + \varepsilon_3 \tag{12}$$

where $\varepsilon_3 = \varepsilon_1 + \varepsilon_0$ and ε_3 and ε_2 are bivariate normal, etc., with covariance $\sigma_{23} = \sigma_{12}$. (Note that equations (12) and (3) are the same except for assumptions about ε_3.) Turning next to the selection equation, we see that Z = 1 is equivalent to $\theta_2' X + \varepsilon_2 > 0$ which in turn implies $\varepsilon_2 > -\theta_2' X$ and $\varepsilon_2/\sigma_2 > -\theta' X$ where $\theta' = \theta_2'/\sigma_2$. But (ε_2/σ_2) is a standard normal variable independent of X. And since Z is binary it follows that

$$E(Z|X) = \mathrm{Prob}(Z=1|X) = 1 - F(-\theta' X) = F(\theta' X) \; , \tag{13}$$

where $F(\cdot)$ is the standard normal cummulative distribution function. Furthermore,

$$E((\varepsilon_2/\sigma_2)|X,Z=1) = f(\theta' X)/F(\theta' X) \tag{14a}$$

and

$$E((\varepsilon_2/\sigma_2)|X,Z=0) \equiv f(\theta' X)/(1 - F(\theta' X)) \; , \tag{14b}$$

where $f(\cdot)$ denotes the standard normal density function. Equations (14a) and (14b) can be rewritten in combined form and rearranged to give

$$\begin{aligned} E((\varepsilon_2/\sigma_2) &= \frac{f(\theta' X)(Z - F(\theta' X)}{(1 - F(\theta' X))F(\theta' X)} \\ &= h(X,Z;\theta) \end{aligned} \tag{15}$$

or, equivalently,

$$E(\varepsilon_2|X,Z) = \sigma_2 h(X,Z;\theta) \quad .$$

Also,

$$E(\varepsilon_3|X,Z) = (\sigma_{12}/\sigma_2^2)E(\varepsilon_2|X,Z) = (\sigma_{12}/\sigma_2)h(X,Z,\theta) \quad . \tag{16}$$

Given (16), the expectation of (12) conditional on X and Z is then

$$E(Y|X,Z) = \theta_1'X + \alpha Z + (\sigma_{12}/\sigma_2)h(X,Z;\theta) \quad . \tag{17}$$

Equation (17) is the conditional expectation function relating observable values and its parameters (θ_1, α_1,σ_{12}/σ_2, $\theta = \theta_2/\sigma_2$) can be estimated by non-linear least squares. The crucial feature of this expression is the inclusion of $h(X,Z;\theta)$ which takes the conditional relationship between X and Z into account, thus removing a source of bias (omission of a variable) in estimating α, the program effect.

In practice (17) is estimated by a two-step procedure (Heckman, 1976) whereby θ ($=\theta_2/\sigma_2$) is estimated by maximum-likelihood probit analysis of Z on X, these estimates are inserted in (15) to estimate $\hat{h} = h(X,Z;\theta)$ for each observation, and then θ_1, α, and (σ_{12}/σ_2) are estimated by linear least-squares regression of Y on X, Z, and $\hat{h}$. There is an alternative estimation procedure attributed to Maddala and Lee (1976) that operates in a similar fashion.

The essential feature of the Heckman-Maddala-Lee procedures is that they resolve the problem of selectivity bias by modifying the outcome equation for presumed selection process effects. As in simple ANCOVA, the adjustment is only necessary in those conditions where treatment selection (Z) and true ability (W) are related after controlling for the observed covariates (X). Thus, if there is no relationship between ε_1 and ε_2 ($\sigma_{12} = 0$), then no bias is introduced through selection, and the more complicated selection modeling adjustments are unnecessary.

In their review, Barnow et al. (1980) cite a number of problems with the selection modeling that require further attention:

(1) which consistent estimation procedure is best,

(2) how to deal with severe collinearity in the second-step regression,

(3) the effect of non-norma' disturbances on the robustness of estimators,

(4) misspecification of the original model, and

(5) multiple selection rules.

Several of these problems have since been addressed to some degree (e.g., see Goldberger, 1980; Heckman, 1980; and Olsen, 1979 on the effects of the departures from normality).

Our reading of the current view (Muthen (1981) is the most recent and comprehensive we have seen) is that the consequences are quite serious (i.e., the procedures fail to remove the selectivity bias) when errors in the regression relation depart from normality and/or homoscedasticity (e.g., Goldberger, 1980; Hurd, 1979; Olsen, 1979) and when the functional form of the selection and/or outcome relations are misspecified. The latter can take several forms. For example, it may be that the true relationship of program and ability to outcome is nonlinear though the specification includes only linear effects. Such a situation might suggest the need for adjustments via selection modeling when a more appropriate modification requires a shift to a new functional form for the relationships.

The second form of specification problem that is likely to occur quite frequently is when relevant variables are omitted from the selectivity bias adjustment. In the Heckman procedures, this problem is manifested by leaving out variables that should be incorporated in the probit step. Again, the consequence is the failure to properly adjust estimates in the outcome equation (Muthen, 1981 reviewing work (not currently available for citation) by Cronbach and Goldberger).

Two other concerns raised earlier about otner approaches to analysis of quasi-experimental data warrant mention here. First, virtually all of the econometric discussions of selection modeling ocus on a single outcome measure. Second, the possibility of measurement errors associated with any of the observable variables (either Y's or X's) is not discussed.

Surely one would want to be able to deal with multiple outcomes and with latent exogeneous (explanatory) variables. At the least it would be helpful to state the expressions for selection and outcome modeling in terms of latent, rather than fallible observed variables. Work by Muthen, Joreskog, and Sorbom (Muthen & Joreskog, 1981; Sorbom, 1978, 1981; Sorbom & Joreskog, 1981) represent initial attempts at selection modeling with latent exogenous variables. Essentially one first estimates latent variables via LISREL and then applies the Heckman procedures using the latent variables rather than the observed set of X's. Unfortunately, these methods of estimating latent variables are currently restricted to models with strictly continuous X variables because of their reliance on maximum likelihood procedures that require multivariate normality.

The above concerns notwithstanding, the selection modeling procedures developed by economists clearly offer improvements over the ANCOVA methods described earlier. Though the demands for careful thinking about selection mechanisms are severe, the rewards of such efforts are often substantial, both analytically and substantively.

Summary. We have described in some detail both the basis for concerns about bias in quasi-experimental studies and two sets of analytical developments (the value-added approach and selection modeling) intended to remove or adjust for bias. Both procedures are improvements over the past mainly because they employ explicit models of the phenomena believed to be responsible for the difficulties in estimating program effects. Both approaches are also adaptable to situations where there are no specific comparison or control groups (instead the effects of specific program features are to be estimated) and where panel data exists on program participants.

Neither approach directly addresses such concerns as measurement errors in the explanatory variables, changes in the scales of measurement over time and changes in the structure of behavior over time. Multiple measures of both exogenous and endogeneous variables with known scale properties are needed to gain a better grip on these problems. If these problems can be alleviated, selection and growth modelinq can become even more widely useful.

Structural Equation Modeling

At various points in the discussions of improvements in analyses of non-equivalent control group designs, we encountered lingering concerns about the nature of the model specification for both selection processes and outcomes, fallible measurements, the handling of multiple indicators, changing scales of measurement and changes in the structure of behavior over time. Resolution of the first of these concerns is never complete; one progresses through obtaining better understanding of the phenomena under investigation (both its elements (constructs) and their interrelations). "Better" theories are the only answer. The combination of improvements in the accumulated wisdom on given phenomena (i.e., better thinking about how a program works and about its possible consequences) and better operationalization of the elements of one's theoretical model (i.e., more comprehensive and valid measurement of its constructs) are a necessary foundation for positive increments in the quality of investigations of social programs. Analytical methods for handling the remaining concerns cited in the opening paragraph of this section (namely fallible measurements, multiple indicator, changing scales of measurement and structure of behavior over time) would seem to be useful to ensure that better thinking and operationalization is reflected in better data analysis and interpretation. Such analytical advances would seem to be particularly pertinent to the broad conception of large-scale program evaluation advocated here.

In theory, the techniques of structural equation modeling with latent variables (see Bentler, 1980; Bentler and Woodward, 1979; Bilby and Hauser, 1979; Goldberger and Duncan, 1973; Joreskog, 1980, 1973, 1974, 1977; Joreskog and Sorbom, 1976, 1978; Sorbom and Joreskog, 1981; Wiley, 1973) appear to be particularly well-suited for resolving several of the remaining methodological problems cited above. These techniques are designed to estimate the unknown coefficients in specified "causal" structures among latent (unobservable) variables.[4] The references cited above provide extensive discussions of the current state of work on structural equation modeling including indications of the kinds of substantive and methodological problems for which these techniques are applicable. Most of the literature addresses mainstream social research issues. However, there have been several applications in educational research contexts (see Lomax (1981) for partial bibliography of educational research applications; however, one of the most comprehensive and carefully documented applications of these methods to educational questions (namely, Munck, 1979) and recent applications with hierarchical data (Keesling, 1978; Wisenbaker, 1980; Wisenbaker and Schmidt, 1978) are not cited).

Existing applications in large-scale educational evaluations are even more limited. The best known is the exchange between Magidson (1977, 1978) and Bentler and Woodward (1978, 1979) on the effects of Head Start. Abt and

Madigson (1980) also use structural equation modeling in their evaluation of a specific school reform. Sorbom and Joreskog (1981) discuss how these techniques can be applied in evaluation research. Finally, structural equation modeling of latent variables is the primary analytical method in the longitudinal examinations of the effects of the characteristics of the educational process and students' background on academic achievement during elementary school years [conducted as part of System Development Corporation's (SDC) Sustaining Effects Study; see Wingard, 1980] and was one of the analytical methods used in SDC's cross-sectional study of the effects of instruction on the achievement growth of compensatory-education students (Wang, et. al., 1981). Given the prominence (and cost) of the Sustaining Effects Study among the set of recent large-scale evaluations in education, we are likely to see additional attempts to apply these methods, assuming of course the continuation of large-scale qualitatively oriented evaluations.

We will not attempt to recount in detail the various analytical nuances of structural equations modeling with latent variables. Instead the general strategy employed by Joreskog and his associates in their LISREL (Linear Structural, Relations) modeling will be described. We then provide a partial accounting of the specific analytical problems in program evaluations that can be addressed, at least in part, by these methods. As with the analytical developments considered earlier, we conclude with a discussion of what we perceive to be the main limitations of structural equation modeling in evaluation contexts.

Basic approach. In currently available variants of structural equation modeling, one begins with a theoretical model about the structural (perhaps causal) relations among a set of pertinent latent (unobservable) constructs (e.g., student background and ability, program and instructional quality, schooling context, student performance). One attempts to operationalize these constructs through the collection of information on observable indicators of each construct (say, measures of aptitudes and some quality at time of program entry; measures of program and instructional characteristics (e.g., emphasis, intensity); measures of environmental characteristics (ability, composition, perceived climates); measures of cognitive, affective, and social outcomes).

The information from these indicators has an observed covariance structure (i.e., each variable yields observed estimates of variance as well as exhibiting covariation with other observed variables). One then estimates the relationships among latent variables and of latent variables to observed variables via statistical means and attempts to reconstruct the observed variance-covariance structure (matrix of variances and covariances) from the estimated variances and covariances implied by the theoretical specification. At this point one judges the acceptability of the fit of the estimated structure to the observed structure, and depending on one's perspective (there is lots of debate about what to do next), either stops or goes through another iteration of the specification-estimation process if the results are unsatisfactory.

LISREL. As we said earlier, the LISREL model developed by Joreskog and associates associates (Joreskog, 1973, 1974, 1977; Joreskog and Sorbom, 1978) is the most widely used analytical approach to estimation in structural equation modeling. This method handles a set of linear structural relations. "The variables in the equations system may be latent variables and there may be multiple indicators or causes of each latent variable...the method allows for both errors in equations (residuals, disturbances) and errors in the observed variables (errors of measurement, observational errors)...yields estimates of the residual covariance matrix and the measurement error covariance matrix as well as estimates of the unknown coefficients in the structural equations, provided that all these parameters are known (Joreskog, 1980, p. 106)"

There are two submodels in the LISREL estimation of structural relations among latent variables. There is a structural model which specifies the relationship among latent variables. In addition, there is a measurement model which specifies the relationships of the measured variables to the unobserved constructs. Typically, there are multiple indicators of each latent construct. The interrelationships among the observed indicators of the same construct are then used to separate the presumed underlying true constructs from the irrelevant and error components of each measure.

The analyst starts with a specification of the structural model and the measurement model. If the unknown parameters in both parts of the model are identified (i.e., there are at least as many observed variances and covariances as parameters to estimate) and if the measured variables have a multivariate normal distribution, maximum-likelihood estimates for the parameters are provided along with accompanying standard errors. There are also procedures for testing lack of fit for all or part of the model (e.g., Bentler and Bonnett, 1981). More formally, the LISREL model can be specified as follows. Let $\underset{\sim}{\eta} = (\eta_1, \eta_2, \ldots \eta$ and $\xi = (\xi_1, \xi_2 \ldots \xi_m)$ be random vectors of latent dependent (endogenous) variables and independent (exogeneous) variables. In a simple input-process-outcome model of program impact with non-experimental data, the latent variables in ε might be socioeconomic background (ξ_1) quality of the home (ξ_2) and student ability (ξ_2). The latent dependent variables would be program quality (η_1; program quality is treated as endogenous because it is viewed as determined in part by the specific input characteristics of students) and program outcomes such as cognitive (η_2) and social (η_3) functioning. The system of linear structural relations is given by

$$\mathbf{B}\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \ , \qquad (18)$$

where $\mathbf{B}$ and $\boldsymbol{\Gamma}$ are coefficient matrices for the relations among endogenous variables (e.g., between η_1 and η_2) and of the exogeneous variables to the endogeneous varaiable (e.g., ξ_2 to η_2) and $\boldsymbol{\zeta}$ is a random vector of residuals (errors in equation, random disturbance terms).

The vectors $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are not observed. Instead we observe vectors $\mathbf{Y} = (Y_1 \ldots, Y_p)$ and $\mathbf{X} = (X_1 \ldots X_q)$ which are indicators of the latent endogeneous and exogeneous variables, respectively. For example, program quality (η_1) might be measured by the opportunity to learn relevant curriculum (Y_1) and the quality of the presentation of the material (Y_2). Cognitive functioning (η_2) might be measured by reading (Y_3) and mathematics achievement tests (Y_4) and social functioning by sociometric measures of friendship networks (Y_5), and teacher ratings of social functioning (Y_6). Observed indicators of the latent exogeneous variables might be family income (X_1) and mother and father's education (X_2 and X_3) for socioeconomic background (ξ_1); availability of learning resources (X_4) and parental aspirations for their child (X_5) for quality of the home (ξ_2), and pretests on reading (X_6) and mathematical skills (X_7) for student ability (ξ_3). The system of equations expressing the measurement model can be written as

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon} \ , \\ \mathbf{x} &= \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta} \ , \end{aligned} \qquad (19)$$

where $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_x$ are matrices of regression coefficients relating $\boldsymbol{\eta}$ to $\mathbf{y}$ and $\boldsymbol{\xi}$ to $\mathbf{x}$, respectively and $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$ are vectors of errors of measurement in $\mathbf{y}$ and $\mathbf{x}$, respectively.

(3) Measuring changes in the scaling of variables over time (e.g., Joreskog, 1979, Sorbom, 1979a).

(4) Detecting changes in the structure of behavior over time (Joreskog, 1979; Shavelson, Bolus and Keesling, 1981).

(5) Detecting differences in the structural relations across groups (e.g., Bentler and Woodward, 1978; Sorbom, 1979b, 1979c).

The first four applications select contributions targeted toward specific concerns that arise in quasi-experimental and non-experimental evaluation studies. The last application allows analysts to compare specific program alterpatives (e.g., participation in Title I vs. Follow Through or High Scope vs. Direct Instruction Follow Through Models, e.c.,) in a more sensitive, comprehensive, and, we believe, sensible way.

Limitations. Unfortunately, as with most analytical advances, there are important practical limitations in applying structural equation modeling in general and LISREL, specifically. The most serious and endemic problem is that the adequacy of the methods is inherently dependent on the quality of the model specification--both the limits of current theory (which constructs arc pertinent) and of current operationalization through the measures one collects. Bad theory and bad data are no less bad simply because we analyze them in a sophisticated and complicated fashion. It is unclear whether the consequences of these shortcomings are more severe in structural equation models though the appearance of sophistication whenever parsimonious and simple examinations are flawed would seem to be a dangerous attribute of any analytical technique.

Another potentially serious limitation is the question of robustness of LISREL to violation of multivariate normality assumptions. Current versions of LISREL are not well-suited for such complications of discrete

If Σ represents the population covariance matrix among the p and q measured variables (13 in our hypothetical example, 6 indicators of endogeneous variables and 7 of exogeneous variables), the elements of this matrix can be expressed as functions of the elements of the four matrices of regression parametrics (Λ_y,Λ_x,B,Γ), the covariance matrix among the exogeneous latent variables ε (typically denoted by ϕ), and the covariance matrices of the errors in the struvtural (ψ) and measurement (θ_ε and θ_δ) models. In application some of these elements are fixed (assigned given values), others are constrained (unknown but equal to one or more other parameters) and the remainder are free parameters to be estimated by the procedures.

Areas of application in evaluation contexts. In most practical applications of LISREL, one focusses on estimating the regression parameter matrices (B,Γ,Λ_y and Λ_x). The ultimate intent is obviously to represent the true structural relationships. The specific analytical problems in program evaluation that LISREL can handle are those that arise in many social research settings. LISREL may be used to deal with a number of problems simultaneously (e.g., Madidson, 1977, Bentler and Woodward, 1978) or may be restricted to handling a single problem (e.g., perhaps obtaining estimates of latent variables for use in selection modeling, or for estimating the factor structure among observable indicators).

Particular applications include:

(1) Correcting for the effects of measurement error (e.g., Keesling and Wiley) in quasi-experiments.

(2) Taking both irrelevance (specific factors unrelated to the construct of interest but present in measured variables) and measurement errors into account (e.g., Linn and Werts, 1977).

measures of independent and dependent variables (except for the multiple group comparison application). Muthen (1979) has worked out procedures for handling certain structural models involving dichotomous variables (e.g., factor analysis of dichotomous variables) but they are not nearly as comprehensive as LISREL. Some researchers have turned to a related set of methods, partial least-squares (PLS), developed by Wold (see McGarvey and Bentler, 1980) because they do not require the multivariate normality. However, in the few empirical examples currently available, the estimates from LISREL and PLS are not very different and the rationale for PLS remains more obscure.

Despite some initial forays by Schmidt and others (Keesling, 1978; Schmidt, 1969; Wisenbaker, 1980; Wisenbaker and Schmidt, 1978), structural equation models for analyzing the hierarchical data frequently encountered in evaluations remain underdeveloped. It is simply too early to tell how to proceed in the area.

Finally, even though the primary reason many investigators turn to LISREL is its ability to estimate complex models with multiple latent constructs and multiple measurements, the practical reality is that LISREL estimation is often overwhelmed by the sheer size and complexity of such models. There are too many ways to go wrong. With large data sets with lots of parameters, practically inconsequential differences in parameters cause statistical fit indices to be significant (necessitating modification of the model). Though LISREL is capable of simultaneously estimating measurement and structural models, in practice researchers with a large number of varaibles often have to estimate these models in separate stages. And the analyses are very expensive by current standards for cost of alternative, though simplified, analytical methods. In his analyses of the SES study of longitudinal data (Wingard (personal communication)) estimates that his typical computer run involving roughly 8 latent constructs with 3 to 10 indicators each costs roughly $250 and often may not even converge to within acceptable limits for the maximum-likelihood estimation.

So, again, we find ourselves with an obvious improvement in analytical methods that is applicable in large-scale program evaluation but is flawed in important respects. Clearly, structural equation modeling is a tool worth having but also one that must be used cautiously.

Concluding Remarks

In our examination of two general classes of analytical methods we have attempted to highlight why they might be considered, how they can be applied, and the limitations on their application. We could have taken each major area of analytical improvements in the past few years and treated them similarly (see, for example, the excellent review of Traub and Wolfe (in press) of the promise and problems in latent trait models for educational measurement).

But this is as it should be. Empirical investigations, be they randomized experiments or simply "passive observational studies", have their imperfections and special shortcomings. Thus, it is not surprising that there is no handy-dandy analytical method that solves all problems. The design and analysis perspective advocated here and presumably shared by Cook (1974, 1981) and Cronbach et. al. (1980), (see also Burstein (1981)) does not require that any one method be without flaws. Instead, it is the weight of the evidence from multiple analyses (and reanalyses) on perhaps overlapping but separatable questions and sets of data that should guide interpretation.

One last caveat. After beginning our work on analytical advances, we quickly became convinced that there were more fundamental problems in the area of data collection in program evaluations that greatly limit the payoff from analytical developments. In fact, we view data collection as the "Achilles Heel" of program evaluation, especially in the way it vitiates the validity of data analysis and interpretation. Elsewhere we (Burstein, Freeman and Sirotnik, 1981) have outlined our reasons for concerns about data collection. At some point, methodologists working in the area of program evaluation will devote greater attention to data collection problems. If not, the next generation of evlauation studies are destined to suffer the fate of the last generation's despite their enhanced analytical power.

Footnotes

1. We simply do not subscribe to the conspirational view of the shift in emphasis (essentially, if you can't find significant effects, change the question) as characterized in several recent accounts of the political history of the evaluation of social programs. Certainly, social programs develop a political constituency (often labeled Stakeholders) consisting of legislators, bureaucrats, service providers, program participants, members of the public as well as evaluators that have a stake in maintaining program activities. These programs also develop enemies (political and ideological) and suffer through internal bickering and lack of common perspective. Yet the interplay of competing forces surrounding any societal activity that has political, economic, and social consequences is the norm rather than the unusual. Moreover, this interplay introduces its own set of dynamics that affect the activity in complex and often unknown ways. Over time a more refined articulation of activities (expected and actual) and their consequences (expected and actual) evolve. It is only natural, then, that the search for better understanding also shifts to more sophisticated and sensitive methods for explicitly linking activities with their consequences.

2. This part of the presentation draws heavily from Barnow et. al. (1980).

3. Tuma and Hannan's work (Tuma and Hannan, 1978; Tuma, Hannan, and Groenveld, 1978) grounds the analysis of changes over time on a categorical dependent variable in a continuous-time stochastic model. They start with a continuous-time Markov model, extend it to deal with population heterogeneity (e.g., differences in background and program characteristics) and time dependence, and develop a maximum-likelihood estimation procedure for estimating the model

from what they call "event-histories" (data giving the number, timing and sequence of changes for a categorical dependent variable). These methods seem to be responsive to certain concerns addressed in the Bryk and Weisberg value-added analysis (i.e., dynamic models of change processes) as well as the econometric selection modeling (dealing with various selection problems such as attrition and systematic selection). However, the techniques are currently restricted to discrete outcome variables (e.g., decision to attend college or not; or college dropout decision) while the present review in restricted to evaluation studies in which the outcomes are viewed as essentially continuous dimensions.

4. We have chosen to use the term "structural equation" modeling rather than the label "causal" modeling more widely used in educational and psychological applications. In our view, the latter term attracts too much criticism about whether phenomena are truly "causal" as opposed to simply relational. This criticism detracts from the analytical potential inherent in these statistical aspects of the models. No one denies that practice in less than ideal (i.e., we never really know the causes in non-experimental studies (or experimental ones for that matter) and this misspecification is an inherent property of empirical social research. Misspecification, in turn, inevitably leads to flawed estimation. Nonetheless, one can conceive of a continuum of better vs. worse empirical approximations to reality. We contend that structural equation modeling with latent variables can potentially yield results that approach the "better" role of the continuum and thus should not be excluded because they are flawed (some philosopher might judge them "wrong".)

REFERENCES

Abt, W.P., & Magidson, J. Reforming schools: Problems in program implementation and evaluation. Beverly Hills, CA: Sage Publications, Inc., 1980.

Anderson, R.B. Follow Through: Testing one model of evaluation and several models of compensation. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, April 1976 (ERIC #120-238).

Barnes, R.E., & Ginsburg, A.L. Relevance of the RMC models for Title I policy concerns. Educational Evaluation and Policy Analysis, 1979, 1(2), 7-14.

Barnow, B.S. The effects of Head Start and socioeconomic status on cognitive development of disadvantaged children. Unpublished doctoral dissertation, University of Wisconsin-Madison, 1975.

Barnow, B.S., Cain, G.G., & Goldberger, A.S. Issues in the analysis of selection bias. In E.W. Stromsdorfer and G. Farkas (Eds.), Evaluation studies review annual, volume 5. Beverly Hills, CA: Sage, 1980.

Bentler, P.M. Multivariate analysis with latent variables: Causal modeling. Annual Review of Psychology, 1980, 31, 419-459.

Bentler, P.M., & Woodward, J.A. Nonexperimental evaluation research: Contributions of causal modeling. In L.E. Datta & R. Perloff (Eds.), Improving evaluations. Beverly Hills, Ca: Sage Publications, 1979.

Boruch, R.F. Bibliography: Randomized field experiments for planning and evaluating social programs. Evaluation, 1974, 2, 83-87.

Boruch, R.F., & Cordray, D.S. An appraisal of educational program evaluations: Federal, state, and local agencies. Washington, D.C.: U.S. Department of Education.

Bryk, A.S. An intvestigation of the effectiveness of alternative adjustment strategies in the analysis of quasi-experimental growth data. Unpublished doctoral dissertation, Harvard Graduate School of Education, 1977.

Bryk, A.S., Strenio, & Weisberg, H.I. A method for estimating treatment effects when individuals are growing. Journal of Educational Statistics, 1980, 5(1), 5-34.

Bryk, A.S., & Weisberg, HI.I Use of nonequivalent control group design when subjects are growing. Psychological Bulletin, 1977, 83, 950-962.

Bureau of Education for the Handicapped. Progress toward a free appropriate public education: An interim report.to Congress. Washington, D.C.: Department of Health, Education, and Welfare, 1978.

Burstein, L. Analysis of multilevel data in educational research and evaluation. In D. Berliner (Ed.), Review of Research in Education, Vol. 8. Itasca, IL: F.E. Peacock, 1980. (a)

Burstein, L. The role of levels of analysis in the specification of education effects. In R. Dreeben & J.A. Thomas (Eds.), Educational production: A microanalysis of schooling. New York: Ballinger Press, 1980. (b)

Burstein, L. Investigating social programs when individuals belong to a variety of groups over time: Implications for Follow Through research and evaluation. Paper presented at the National Institute of Education Conference on Follow Through Research and Development, Pittsburgh, PA, March 1981.

Cain, G.C. Regression and selection models to improve nonexperimental comparisons. In C.A. Bennett and A.A. Lumsdain (Eds.), Evaluation and experiment, some critical issues in assessing social programs. New York: Academic Press, 297-317.

Campbell, D.T., & Erlebacher, A.E. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), The disadvantaged child. Vol. 3, New York: Brunner/ Mazel, 1970.

Cicirelli, V.G., et al. The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development. Athens, OH: Ohio University and Westinghouse Learning Corporation, 1969.

Cline, M.D., et al. Education as experimentation: Evaluation of the Follow Through planned variation model (Vols. 1A, 1B). Cambridge, MA: Abt Associates, 1974.

Cohen, D.K., & Garet, M.S. Reforming educational policy with applied social research. Harvard Educational Review, 1975, 45(1), 17-43.

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

Cook, T.D. Dilemmas in evaluation of social programs. In M.B. Brewer and B.E. Collins (Eds.), Scientific inquiry and the social sciences, San francisco, CA: Jossey Bass, Inc., 1981, 257-287.

Cook, T.D., & Campbell, D.T. Quasi-experimentation. Chicago, IL: Rand McNally College Publishing Company, 1979.

Coulson, J.E. National evaluation of the Emergency School Aid Act (ESAA): Review of methodological issues. Journal of Educational Statistics, 1978, 3(1), 1-60.

Coulson, J.E., Ozenne, D.G., Hanes, S.D., Bradford, C., Doherty, W.J., Suck, G.A., & Hemenway, J.A. The third year of Emergency School Aid Act (ESAA) implementation. System Development Corporation, TM-5236/014/00, 1977.

Cronbach, L.J. Design educational evaluations. Stanford Evaluation Consortium, Stanford University, 1978.

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D., Hornik, R.C., Phillips, D.C., Walker, D.F., & Weiner, S.S. Toward reform of program evaluation. San Francisco, CA: Jossey-Bass, Inc., 1980.

Cronbach, L.J., Rogosa, D.R., Floden, R.E., & Price, G.G. Analysis of covariance in nonrandomized experiments: Parameters affecting bias. Occasional Paper, Stanford Evaluation Consortium, Stanford University, 1977.

Cross, C.T. Title I evaluation -- A case study in congressional frustration. Educational Evaluation and Policy Analysis, 1979, 1(2), 15-22.

Daillak, R.H., & Alkin, M.C. Qualitative studies in context: Reflections on the CSE studies of evaluation use. Los Angeles, CA: University of California, Los Angeles, Center for the Study of Evaluation, 1981.

Firebaugh, G.L. Groups as contexts and frog ponds. In K. Roberts and L. Burstein (Eds.), New directions in the methodology of social and behavioral research. San Francisco, CA: Jossey-Bass Publishers, 1980.

Goldberger, A.S. Abnormal selection bias. 9006, Social Systems Research Institute, University of Wisconsin, Madison, 1980.

Goldberger, A.S. Methods for eliminating selection bias. Memorandum, Department of Economics, University of Wisconsin, Madison, 1979.

Goldberger, A.S. Selection bias in evaluating treatment effects: Some formal issutrations. Discussion paper 123-72. Madison: Institute for Research on Poverty, 1972.

Haney, W. A technical history of the national Follow Through evaluation, Vol. V, The Follow Through planned variation experiment. Cambridge, MA: The Huron Institute, 1977.

Heckman, J. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement, 1976, 5, 475-492.

Heckman, J. Addendum to 'sample selection bias as a specification error'. In E.W. Stromsdorter and G. Farkas (Eds.), Evaluation studies review annual, vol. 5. Beverly Hills, CA: Sage Publications, 1980.

House, E. Evaluation with validity. Beverly Hills, CA: Sage Publications, 1980.

House, E. The logic of evaluative argument. CSE Monograph. Los Angeles, CA: CEnter for the Study of Evaluation, 1977.

House, E., Glass, G.V., McLean, L.D., & Walker, D.F. No simple answer: Critique of the 'Follow Through' evaluation. Harvard Educational Review, 1978.

Joreskog, K.G. A general method for analysis of covariance structures. Biometrika, 1970, 57, 239-251.

Joreskog, K.G. A general method for estimating a linear structural equation system. In A.S. Goldberger & O.D. Duncan (Eds.), Structural equation models in the social sciences. New York: Seminar Press, 1973.

Joreskog, K.G. Analyzing psychological data by structural analysis of covariance matrices. In D.H. Krantz, R. Atkins, D.Luce, and P. Supper (Eds.), Contemporary developments in mathematical psychology, Vol. II, San Francisco, CA: D.W. Freeman & Co., 1974.

Joreskog, K.G. Structural equations models in the social sciences: Specification, estimation, and testing. In P.R. Krishnaiah (Ed.), Applications of statistics, Amsterdam: North Holland Publishers, 1977.

Joreskog, K.G., & Sorbom, D. Statistical models and methods for analysis of longitudinal data. In D.J. Aigner & A.S. Goldberger (Eds.), Latent variables in socioeconomic models, Amsterdam: North Holland Publishers, 1977.

Joreskog, K.G., & Sorbom, D. Advances in factor analysis and structural equation models. Cambridge, MA: Abt Books, 1979.

Klibanoff, L.S., & Haggart, S.A. Report #8: Summer growth and the effectiveness of summer school. Technical Report #8 from the Study of the Sustaining Effects of Compensatory Education on Basic Skills, System Development Corporation, Santa Monica, CA, 1981

Maddahian, E. Statistical models for the study of cognitive growth. Unpublished doctoral dissertation, University of California, Los Angeles, 1981.

Maddala, G.S., & Lee, L.F. Recursive models with qualitative endogenous variables. Annals of Economic and Social Measurement, 1976, 5, 525-545.

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

Munck, I.M.E. Model building in comparative education. Stockholm, Sweden: Almqvist & Wiksell International, 1979.

Muthen, B. Some categorical response models with continuous latent variables. In K.G. Joreskog & H. Wold (Eds.), Systems under indirect observation: Causality, structure and prediction. Amsterdam: North Holland Publishing Company, 1981.

Muthen, B., & Joreskog, K.G. Selectivity problems in quasi-experimental studies. Presented at the Conference on Experimental Research in Social Sciences, University of Florida, Gainesville, 1981.

National Institute of Fducation. Evaluating compensatory education. A report on the NIE ompensatory Education Study, 1977.

Olsen, R.J. Tests for thepresence of selectivity bias and their relation to specifications of functional form and error distribution. Working paper No. 812, Yale University, 1979.

Raizen, S.A., & Rossi, P.H. (Eds.) Program evaluation in education: When? How? To what ends? Washington, D.C.: National Academy Press, 1981.Reichardt, C.S. The statistical analysis of data from non-equivalent group designs. In T.D. Cook and D.T. Campbell (Eds.), Quasi-experimentation, Chicago: Rand McNally, 1979.

Rogosa, D. Politics, process, and pyramids. Journal of Educational Statistics, 1978, 3(1), 79-86.

Rossi, P.H., & Lyall, K.C. Reforming public welfare: A critique of the negative income tax experiments. Russell Sage Foundation, 1976.

Sirotnik, K.A., & Oakes, J. A contextual appraisal system for schools: Medicine or madness? Educational Leadership, 1981 (in press).

Stallings, J.A., & Kaskowitz, D.H. Follow Through classroom observation evaluation 1972-1973. Stanford Research Institute, August 1974.

Smith, M.S., & Bissell, J.S. Report analysis: The impact of Head Start. Harvard Educational Review, 1970, 40, 41-104.

Strenio, J.F., Bryk, A.S., & Weisberg, H.I. An individual growth model perspective for evaluation of educational programs. Proceedings of the social science section, Annual Meetingof the American Statistical Association, 1977.

Stromsdorfer, E.W., & Farkas, G. Evaluation studies review annual, Vol. 5. Beverly Hills, CA: Sage Publications, 1980.

Sorbom, D. An alternative to the methodology for analysis of covariance. Psychometrika, 1978, 43, 381-396.

Sorbom, D. Structural equation models with structured means. To appear in K.G. Joreskog and H. Wold (Eds.), Systems under indirect observation: Causality, structure, prediction. Amsterdam: North Holland Publishing Co., 1981.

Sorbom, D., & Joreskog, K.G. The use of structural equation models in evaluation research. Presented at the Conference on Experimental Research in Social Sciences, University of Florida, Gainesville, 1981.

Tuma, N.B., Hannan, M.T., & Goroenveld, L. Dyanmic analysis of event histories. In E.W. Stromsdorfer, & G. Farkas, Evaluation studies review annual, Vol. 5, Beverly Hills, CA: Sage Publications, 1980.

Wargo, M.J. An evaluator's perspective. In M.J. Wargo & D.R. Green (Eds.), Achievement testing of disadvantaged and minority students for educational program evaluation. McGraw-Hill, 1977.

Tuma, N.B., & Hannan, M.T. Approaches to the censoring problem in analysis of event histories. In, K. Schuessler (Ed.), Sociological Methodology, San Francisco: Jossey-Bass, 1979.

Webb, N.M. Group process: The key ot 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.

Wisler, C.E., & Anderson, J.K. Desinging a Title I evaluation system to meet legislative requirements. _Educational Evaluation and Policy Analysis_, 1979, _1_(2), 47-56.