EMPIRICAL STUDIES OF MULTILEVEL APPROACHES TO TEST DEVELOPMENT AND INTERPRETATION

Leigh Burstein

M. David Miller

CSE Report #176
1981

Center for the Study of Evaluation
Graduate School of Education, UCLA
Los Angeles, California 90024

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

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 a 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 participating 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.