DOCUMENT RESUME

ED 244 730 PS 014 289

AUTHOR	Jaeger, Richard M.
TITLE	On the Use of Standardized Achievement Tests in Follow Through Program Evaluation.
INSTITUTION	North Carolina Univ., Greensboro. Center for Educational Research and Evaluation.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Jan 81
CONTRACT	NIE-P-80-0179
NOTE	40p.; Print is light.
PUB TYPE	Viewpoints (120) -- Information Analyses (070)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Achievement Tests; Literature Reviews; Primary Education; *Program Evaluation; Research Needs; *Standardized Tests; *Testing Problems; *Test Use; Test Validity
IDENTIFIERS	Anchor Test Study; Metropolitan Achievement Tests; *Project Follow Through; *Test Equivalence

ABSTRACT

One of several papers commissioned by the National Institute of Education (NIE), this discussion addresses the utility of test-equating studies for large-scale program evaluation and investigates the use of standardized tests in similar evaluation efforts. The discussion begins with an examination of the question, Should standardized achievement tests be used in a programmatic evaluation of Follow Through? Predominant issues in the argument against the use of standardized tests in program evaluation are then addressed. It is concluded that the earlier practice of evaluating Follow Through with a single standardized achievement test should be abandoned. The second question addressed poses the issue of whether NIE should sponsor another large-scale test-equating study in support of Follow Through program evaluation. Consideration of this issue is based largely on experience gained through the administration of the Anchor Test Study (ATS). Following a brief description of ATS history, the discussion draws on an extensive review of the literature to suggest that ATS results have not been considered. Finally, implications of the review for Project Follow Through are presented together with (1) a set of recommendations concerning the use of standardized tests in the evaluation of Follow Through, and (2) recommendations about NIE sponsorship of a major test-equating study involving tests that might be suitable for Follow Through program evaluation. (RH)

ON THE USE OF STANDARDIZED ACHIEVEMENT TESTS IN FOLLOW THROUGH PROGRAM EVALUATION

Richard M. Jaeger
University of North Carolina, Greensboro

CENTER FOR EDUCATIONAL RESEARCH AND EVALUATION
January, 1981

Prepared pursuant to NIE Contract No. NIE-P-80-0179

INTRODUCTION AND OVERVIEW

This paper is one of several commissioned by the National Institute of Education in support of planning for the testing of new approaches in the Follow Through program. It is one of a number of papers that Follow Through planners categorized as "supporting research," intended to consider fundamental methodological and analytic issues that will likely impinge on the planning, design, operation and evaluation of the Follow Through program in the next few years. More specifically, I was asked to address two issues: the utility of test equating for large-scale program evaluation, and the use of standardized tests in large-scale program evaluation.

In responding to the NIE charge, I have chosen to focus my attention more closely on the Follow Through program than the NIE planning staff may have intended. But the issues treated in this paper are clearly pertinent to large-scale evaluations of many instructional programs, especially those intended for educationally disadvantaged students.

The paper begins with an examination of the question: Should standardized achievement tests be used in a programmatic evaluation of Follow Through? There is abundant evidence of divergent views on this issue in the methodological literature. Advocates can be found among both supporters and critics of the longitudinal evaluation of Follow Through that culminated in the ABT Associates reports. Those who advise against the use of standardized achievement tests in the evaluation of any program for disadvantaged students (or, perhaps, in any program evaluation) include such pillars of the measurement community as Ralph Tyler (1972; 1978) and the late Oscar Krisen Buros (1977). Three issues appear to predominate in the argument against the use of standardized tests in program evaluation: (1) the design of the tests makes them insensitive to instruction and more suitable to their original purpose of distinguishing normatively among the achievement levels of individual students; (2) no single

standardized test can be used to validly examine students' achievement of the instructional objectives of the variety of projects, sponsors, and approaches present in virtually all federally-supported education programs. Lack of content congruence makes standardized tests inappropriate as instruments for program evaluation; (3) the political baggage attached to standardized tests results in wholesale overinterpretation of the findings reported for them. Standardized tests have been oversold to the public and to educational policy makers to the degree that they overshadow any alternative measures of educational impact, regardless of their appropriateness or validity. The only prudent course therefore is to avoid the use of standardized tests altogether in large-scale program evaluations. Judgments on these issues are documented and discussed briefly.

Although the first and the third arguments against the use of standardized achievement tests in program evaluation can only be supported by avoiding such use altogether, the second argument, concerning content validity, might be handled responsively by using a variety of standardized achievement tests instead of only one. Presumably, if each sponsor or developer of a project were allowed to select the standardized achievement test that most closely matched the content of his or her instructional model, measurement validity would be materially improved. However, scores on a variety of achievement tests, whether in raw-score or derived-score form can appropriately be aggregated for evaluation of program effects only if the tests have been properly equated.

This leads to the second major question addressed in this paper: Should NIE sponsor another large-scale test-equating study in support of Follow Through program evaluation? Consideration of this issue is based largely in the experience gained through the first large-scale test-equating study supported by the federal government -- the Anchor Test Study, sponsored by the U. S. Office of Education. The history of the Anchor Test Study is reviewed briefly, with particular attention to its intended use as a tool in federal evaluation of the effectiveness of the Title I, ESEA program. Utilization of the results.

of the Anchor Test Study is reviewed next. A thorough search of relevant literature was conducted to determine whether, and to what degree, Anchor Test Study results and data have been used in program evaluation or in measurement and evaluation research. Finally, the implications of this review for the Follow Through program are presented, together with a set of recommendations on the use of standardized tests in the evaluation of Follow Through and on NIE sponsorship of a major test-equating study involving tests that might be suitable for Follow Through program evaluation.

SHOULD STANDARDIZED ACHIEVEMENT TESTS BE USED IN AN EVALUATION OF THE FOLLOW THROUGH PROGRAM?

Arguments against the use of standardized achievement tests in program evaluation are not new. In 1972, when commenting on the suitability of standardized achievement tests for use as assessment devices, Ralph Tyler stated:

> "...the exercises have not been obtained by a systematic sampling of what children are expected to learn. Instead, the exercises comprise a sample of items that differentiate children. It may seem odd . . . but this is due to the fact that most psychometrists since World War I have been interested in individual differences among children and in the process of sorting children rather than in the process of learning."

Tyler expanded on his position in a reaction to a paper by Hoepfner, presented at a USOE-sponsored conference on achievement testing of disadvantaged and minority students for educational program evaluation in 1976 (see Tyler, in Wargo and Green, 1978). Tyler's remarks are central to the question at hand, so it is worthwhile quoting him at length:

> "the confinement of the selection [of achievement tests for program evaluation] to contemporary norm-referenced achievement tests stultifies most of the possibilities for valid and accurate appraisal of the outcomes of educational programs. We can learn very little about the strengths and

weaknesses of programs of 'compensatory education,' or those designed for children of minority groups, from the results of these tests. At best, they are rough and imprecise measures, and most probably they are invalid."

"The first critical assumption made by psychological examiners is that the purpose of the test is to measure individual differences and to arrange those who take the test in a continuum from the best to the poorest. The purpose of program evaluation is to determine how many pupils have learned what the program seeks to teach, and the amount learned, rather than to separate pupils as much as possible. Furthermore, psychological examiners assume that the population of test takers should form a normal distribution similar to the distributions of some of their physical characteristics like height and weight. In contrast, educational programs are designed to help all pupils learn such things as to read, compute, write and understand scientific principles."

Tyler goes on to suggest that the central purposes of the developers of standardized achievement tests leads them to concentrate on items that are near the 50 percent difficulty level, eliminating items that are suitable to the assesement of disadvantaged students and students who are most able. He further suggests that differences between widely used curricula force test makers to sample behaviors that are largely learned out of school, rather than in school, in their attempt to build tests that are universally applicable. Severe problems of invalidity are said to result.

Oscar Krisen Buros, in a reminiscence on fifty years of measurement history, also commented critically on the use of standardized achievement tests in program evaluation. His views mirror those of Tyler:

"[standardized achievement tests are] harmful to the development of the best possible measuring instruments. . .. It seems inescapable that such methods . . . insidiously tend to strengthen the status quo; to

impede curricular progress; to perpetuate our present grade classification; to differentiate rather than to measure, conceal unlearning, and to give an illusory sense of continuous learning from grade to grade" (1977).

Speaking in reaction to a paper by William Coffman at the conference on achievement testing of disadvantaged and minority students referenced above, Jaeger (1978) also advised against the use of standardized achievement tests in evaluating education programs for disadvantaged students:

"The content and skills to be measured by commercially available standardized tests are determined through expert judgments of what typical students in specific grades should know, when exposed to widely used basic skills curriculum materials available for these grades. Are minority and disadvantaged students to be considered typical students? Clearly not, for if they were we would not have specially designed programs to meet their special needs. Are the curriculum materials used in these programs typical of those used with students in these grades? Logic would again tell us that the answer is no. For if standard curriculum materials were used, there would be no need for special programs. So the very process by which standardized achievement tests are planned, if the process follows the ideal, threatens the content validity of these tests for uses that are the subject of this conference. If the content coverage of the tests corresponds to typical curricula, and not to curricula used in the programs to be evaluated, the evaluator may well be judging the status and progress of disadvantaged children on material they have not had the opportunity to learn."

These various judgments on the appropriateness of standardized achievement tests for educational program evaluation are supported by recent research on the effects of the congruence between test content, and curriculum or instructional content, on student achievement. This literature is well reviewed by Tittle (1980), but several relevant

findings are noted here. Jenkins and Pany (1976) conducted a careful analysis of the overlap in vocabulary between five standardized reading achievement tests recommended by their publishers for use in grades one or two, and seven basal reading texts, recommended by their publishers for use in the same grades. The authors estimated the grade equivalent scores that would be earned by students who mastered all of the words in a given reader, and then answered correctly, all of the items on a given test that pertained to words common to the reader and the test. The most extreme variation in scores for a given grade-one reading book was a low grade-equivalent score of 1.0 to a high grade-equivalent score of 2.3. The range of extremes for second-grade books was a low grade-equivalent score below 1.0 if one test was used to a high grade-eqivalent score of 3.4 if another test was used. For the Metropolitan Achievement Word Knowledge Test, grade-eqivalent scores ranged from a low of 1.9 for one reading book to a high of 2.5 for another, at the second-grade level. The range for other tests was typically much greater. In short, the Jenkins and Pany study shows that the content of the achivement test used to evaluate a program is critical to the resulting estimate of its effectiveness. In the evaluation of a program like Follow Through, where a large variety of curricula are purposefully included, some of those curricula must necessarily suffer selection bias if a single test is used for evaluation of student achievement. These conclusions are also consistent with the results of investigations by Armbruster, Stevens & Rosenshine (1977), Chang and Raths (1971), Schutes (1969), Cooley and Leinhardt (1978), Hoepfner (1978) and Bianchini (1978). The latter two studies are discussed in the next section of this paper.

Although federally-supported education programs typically have broad lists of objectives that include alleviation of social, economic and educational deprivation, judgments of the success or failure of these programs have often been grounded in students' performances on standardized achievement tests. The Follow Through program is a

case in point. Schiller, Stalford, Rudner, Kocher and Lasnick (1980) describe Follow Through as a "Federal educational assistance program designed to provide comprehensive services to children from low income families and to increase understanding about effective practices in educating these children" (p.2). In elaborating on the meaning of "comprehensive services," they include health, social, and other support services, in addition to educational services.

Schiller, et al. summarize the principal findings of a nine-year evaluation of the national Follow Through program as follows:

> "There was more variability in outcomes within models from site to site than there was between models;
>
> Models that emphasized basic skills produced more gains in those areas and in self concept than other models;
>
> Overall, there was little difference observed in the performance of Follow Through and non-Follow Through children. Both groups of youngsters remained substantially below national norms"(p.3).

Each of these findings concerns students' performances on pencil-and-paper assessment instruments, and most refer to their performances on the Metropolitan Achievement Tests. These judgments of the merits of the Follow Through program depend on a narrow range of outcome measures, with standardized achievement tests heading the list of those measures. It should be noted that Schiller, et al. accurately reflect the emphases given to various Follow Through outcome measures in the Abt Associates reports on the longitudinal evaluation (Cline, Ames, Anderson, Bales, Ferb, Jorni, Kane, Larson, Park, Proper, Stebbins & Stern,1974) to (Stebbins, St Pierre, Proper, Anderson & Cerva, 1977), as well as in subsequent discussions of those reports in the educational research literature (House, Glass, McLean & Walker, 1978; Anderson, St Pierre, Proper & Stebbins, 1978; Wisler, Burns & Iwamoto, 1978).

It is interesting to note that all of to the papers concerned with the accuracy of

the longitudinal evaluation of the Follow Through program that were presented in the May, 1978 issue of the Harvard Educational Review endorse the use of the Metropolitan Achievement Test as an outcome measure. House, et al. protested the labeling of that test as a "basic skills" measure, suggesting that it assessed a far narrower range of skills, which they labeled the "mechanics of reading and arithmetic." However, they did not protest the use of a standardized achievement test in the national Follow Through evaluation, except on the grounds of insufficiency:

> "The coverage of outcome domains is so poor that no judgment of best model can legitimately be made, no matter how large the difference in test scores" (p.156).

And also on page 156:

> "Even if dependable differences were found on the MAT, such differences would be inadequate evidence of which model is best. Follow Through was to be an investigation of models of comprehensive early childhood education -- not just reading, not just arithmetic, not just language usage. An attempt was made to measure more than a few narrow scholastic outcomes but that attempt was not successful. It serves no one well to proceed as if it had been. Although who did best on the MAT might be a valid question, it would be wrong to confuse that question with the one that was actually asked."

In their response to the House, et al. criticisms, the principal Abt Associates evaluators of the Follow Through program (Anderson, et al., 1978) cite the main conclusions of their report, which they claim their critics ignored. One citation is paticularly telling, in that it illustrates the primacy of the Metrpolitan Achievement Tests in the Follow Through evaluation, and the way in which the language of evaluative reporting can go well beyond the limited scope of the data collected. This citation fuels the argument of those who suggest that standardized achievement tests not be used in

large-scale program evaluations because their results will be overinterpreted. Although they are basing their conclusions solely on MAT scores, Anderson, et al. state:

> "With few exceptions, Follow Through groups were still scoring substantially below grade level at the end of three or four years' intervention (Bock, Stebbins & Proper, 1977, passim). Poor children still tend to perform poorly in school even after the best and the brightest theorists -- with the help of parents, local educators, and federal funds, and supported by the full range of supplementary services associated with community action programs -- have done their best to change the situation."

In their reaction to the House, et al. criticism of the Follow Through evaluation, the USOE program officers who supervised the Abt Associates work (Wisler, et al., 1978) make an interesting claim to the validity of the MAT as a measure of Follow Through effectiveness: "We agree that many model-specific objectives were not measured in the national evaluation, but the goal was to gather valid data on a common set of outcomes generally considered important."

The validity arguments made by Tyler, and Buros were apparently ignored by the contributors to the Harvard Educational Review papers. Although the latter authors debate the adequacy of the MAT as a measure of Follow Through effectiveness, none of them seems to consider the possibility that it was selectively biased against some or most of the Follow Through models. Both sets of defenders of the evaluation base their case on the importance of the content measured by the MAT, and ignore the possibility that that content might not have been a part of the curricula of the Follow Through projects that were evaluated.

Wisler, et al.'s validity claim for the MAT embodies a misuse of the term. The data collected in the Follow Through evaluation are neither valid nor invalid. It is the conclusions and inferences put forth on the basis of those data that must be examined for validity. What appears to be invalid is the conclusion that the Follow Through program

failed just because the MAT scores were low.

In the final analysis, the use of standardized achievement tests in large-scale program evaluation, and in particular, in the Follow Through evaluation, is a matter of judgment. The positions of those who oppose such uses of standardized tests is increasingly supported by evidence on the differential content validity of widely-used tests when applied to early primary level programs in the basic skills. On the basis of that evidence, I would recommend that the earlier practice of selecting a single standardized achievement test for overall evaluation of the Follow Through program be abandoned.

Whether a number of standardized achievement tests should be used in a national Follow Through evaluation, after they have been properly equated, is a separate question that is examined in the balance of this paper.

SHOULD FOLLOW THROUGH SPONSOR ANOTHER MAJOR TEST-EQUATING STUDY?

If data collected using several standardized achievement tests were to be aggregated in a national evaluation of the Follow Through program, it would first be necessary to equate the tests used in the evaluation. Such an undertaking would involve a major investment of funds and time, and might not be feasible economically or technically. The merits of a test-equating study as a tool for Follow Through evaluation can perhaps best be explored by considering in detail the objectives, results, and utilization of the first large-scale test-equating study supported by the federal government, the USOE-sponsored Anchor Test Study. The next section of this paper contains a brief review of the history of the Anchor Test Study, and a detailed review of the utilization of Anchor Test Study results in program evaluation and in measurement and evaluation research. These reviews are used as the basis of recommendations on the advisability of NIE sponsorship of a test-equating project in support of Follow Through evaluation.

THE ANCHOR TEST STUDY
HISTORY, UTILIZATION, AND IMPLICATIONS FOR FOLLOW THROUGH

A Brief History

In the late 1960's, the U.S. Office of Education(USOE) conducted several large-scale surveys for the purpose of securing information that would be useful in judging the operation and impact of Title I of the Elementary and Secondary Education Act of 1965(ESEA). Students' performance on standardized achievement tests, particularly in reading and mathematics, was adopted as a primary indicator of the direction of program services to students who were most educationally disadvantaged. In addition, changes in performance on standardized achievement tests, from the beginning of a school year to the end, were to be used as a primary indicator of programmatic success or failure in alleviating the effects of economic deprivation and educational disadvantagement.

At the time these surveys were initiated, state departments of public instruction and many large school systems resisted the collection of any uniform achievement test data by the U. S. Office of Education. The adverse political impact of the Survey on Equality of Educational Opportunity (Coleman, et al., 1966) was still being felt, and most Chief State School Officers were wary of ay data that would permit the comparison of students' achievement test performances in different states. As a result of this apprehension, it was decided that use of a common achievement test in Title I evaluation surveys was politically infeasible.

The first USOE Survey on Compensatory Education (1968) requested that school systems provide achievement test scores in reading and mathematics, from their existing records, for students in grades two, four, and six, at the beginning and end of the 1968 school year. The reported scores for tens of thousands of students included perofrmances on more than 400 combinations of tests,levels, and forms, administered in the fall and spring of

1968. However, seven major achievement tests accounted for about 90 percent of the scores reported in that year.

The achievement test scores reported by school systems were essentially useless for purposes of Title I evaluation. Once the data were obtained, it was quickly realized that no single publisher's norms were used with a sample of students that was remotely representative of the population being served by Title I,ESEA. The temptation to convert scores on different tests to a common derived scale -- such as grade equivalent scores or T-scores or percentile ranks -- was ignored when analysis of the content of the tests and the publishers' norms revealed substantial differences along both dimensions. One might argue that the Metropolitan Achievement Tests and the Iowa Tests of Basic Skills, say, assess reading comprehension using somewhat similar exercises, and that scores on both tests would be highly correlated were both to be administered to a large randomly-selected sample of fourth-graders. However, the sampling methods used by the publishers of these tests in developing their national norms were appreciably different, as were the cooperation rates of different types of school systems invited to participate in their test normings. As a result, the "national" norms reported by these publishers could not be considered equivalent. And derived scores on these two tests, or on any others, could not legitimately be aggregated for purposes of Title I evaluation.

In 1969, the Bureau of Elementary and Secondary Education in USOE supported a study of the feasibility of equating scores on the reading comprehension and vocabulary subtests of five major test batteries. Expert judges developed a content classification system for the tests and assigned each item on every test to a content category, in an attempt to estimate the congruence of the tests. In addition, triples of tests were administered to several thousand students in the District of Columbia and the states surrounding Washington, D. C. so that correlations among corresponding subtests could be estimated. The judges' attempts to estimate the content similarities of corresponding subtests were not successful. Their lack of agreement on the categorization of items from

a given subtest was so great that their estimates of the congruence of different subtests were suspect. However, the correlations among corresponding subtests in different test batteries were sufficiently high (at least in the high 80's when disattenuated) that a major test-equating study was judged to be feasible.

Early in 1971, Educational Testing Service was awarded a $700,000 contract to conduct an equating and restandardization study involving the seven most widely used reading comprehension and vocabulary subtests in grades four,five, and six. The project came to be known as the "Anchor Test Study" because of the equating methodology employed. It involved restandardization of the reading comprehension and vocabulary subtests of the Metropolitan Achievement Tests intended for use with students in grades four,five and six, using a carefully selected stratified sample of public and non-public elementary schools chosen from counties throughout the United States. More than 200,000 students in these grades provided useable data for restandardization of these subtests. A second part of the Anchor Test Study was an equating study that produced tables of score correspondence between the reading comprehension and vocabulary subtests of the California Achievement Tests, the Comprehensive Tests of Basic Skills, the Iowa Tests of Basic Skills, the Metropolitan Achievment Tests, the Sequential Tests of Educational Progress, the SRA Achievement Series, the Stanford Reading Tests, and in a supplementary study, the Gates-MacGinitie Reading Series. Nearly 135,000 students provided useable test scores for the equating portion of the study conducted in the Spring of 1972, and an additional 14,000 students provided useable test scores the following spring for the supplementary equating of the Gates MacGinitie test to the other seven tests.

Additional details on the design of the Anchor Test Study and its results can be found in the thirty-volume final report on the project (Loret, Seder, Bianchini and Vale, 1972), in the three-volume supplementary report (Loret, Seder, Bianchini and Vale, 1973), and in review articles by Linn (1975) and by Jaeger (1973).

Although the methodological soundness of the Anchor Test Study is uncontestable, one could probably find a variety of views on its ultimate value. Its direct federal cost of three-quarters of a million dollars pales in comparison to the fifty million dollar federal expenditure for evaluation of the Follow Through program. However, its objectives were far narrower. And its findings created far less excitement and controversy in public and policy circles.

The Anchor Test Study clearly established the feasibility of equating the reading comprehension and vocabulary subtests of different test batteries, even though they were not designed to be psychometrically parallel. Prior to completion of the study, it was not certain that the tests to be equated were similar enough make equating possible. In theory, parallel tests can be equated and non-parallel tests cannot (Angoff, 1971). Tests that differ in difficulty, length, reliability, and the constructs they assess will not generally exhibit a consistent relationship across samples of different composition. To be equatable, it is often suggested that a pair of tests have a disattenuated intercorrelation of at least 0.95, a value that is similar to the inter-form correlations of many achievement tests used in the elementary grades.

Correlations among the subtests equated in the Anchor Test Study were typically in excess of the 0.95 criterion. Of 189 correlations between pairs of subtests equated at levels appropriate to students in grades four, five, and six, 106 (56 percent) were in the range 0.98 to 1.00; 53 (28 percent) were in the range 0.95 to 0.97; 26 (14 percent) were in the range 0.92 to 0.94; and only 4 (2 percent) were in the range 0.89 to 0.91. Thus 84 percent of the subtest correlations met the admittedly arbitrary criterion of 0.95. In addition, the standard errors of equating achieved through the Anchor Test Study were consistently less than one-half of a raw-score point at all score levels above the chance scores on the subtests being equated (Loret, Seder, Bianchini and Vale, 1972). Equating errors of this magnitude are generally smaller than test publishers have realized when they equated alternate forms of their tests that were designed to be

parallel.

Utilization of Anchor Test Study Findings

Use in Program Evaluation. Although the Anchor Test Study was intended primarily to provide a tool for use in evaluating Title I, ESEA at the national level (and perhaps at state levels as well), there is little evidence that it has been used for this purpose. In 1979, Stonehill and Fishbein presented a methodological paper on the aggregation of achievement gains in Title I evaluations that referenced data on the comparability of achievement test results produced through the Anchor Test Study. They concluded that the normal curve equivalent scale developed as a part of the Title I Evaluation and Reporting System did not reflect a common score metric, and thus did not produce equivalent achievement scores for students administered different tests. Neglecting this judgment, USOE fostered the adoption of regulations on local Title I evaluation that incorporate the models recommended in the Title I Evaluation and Reporting System.

The conclusions advanced by Stonehill and Fishbein are based in part on a paper by Jaeger (1979) that examined the consistency of achievement gains in the normal curve equivalent metric that would be realized using various reading achievmeent tests equated in the Anchor Test Study. Jaeger concluded that the aggregation of achievement test results in the NCE metric would incorportate measurement errors that were likely to exceed the true gains typically found in Title I evaluations.

Vale and Bianchini (1973) used Anchor Test Study data in completing an analysis of the policy implications of various distributions of federal funds to school systems. In particular, they provided a basis for establishing eligibility criteria for participation in the then-proposed Better Schools Act, using relationships between students' performances on Anchor Test Study tests, certain family background variables (such as parental income category) and certain school system variables (such as degree of

urbanism). Although this application of Anchor Test Study findings is not strictly evaluative, it does fit within the broad framework of federal education program analysis.

In 1974, Jaeger presented a paper on the use of Anchor Test Study results in federal and statewide evaluation of Title I at the Annual Meeting of the American Educational Research Association. The paper enumerated some methodological possibilities, but was based more on conjecture and fond hope than on experience. Its subsequent impact on federal Title I evaluation is not demonstrable.

Apart from these four papers, a thorough serach of the ERIC data base on such descriptors as evaluation methods, federal programs, equated scores, compensatory education programs, achievement tests, and achievement gains failed to produce ay evidence that the Anchor Test Study has been used either by the federal government or by state agencies in the evaluation of any federally-supported education program. If the Anchor Test Study has had any significant impact on evaluation or measurement practice or theory, it is clearly apart from its direct use as a tool in program evaluation.

Use in Educational Research and Assessment. Results from the Anchor Test Study have been used in a variety of educational research and assessment projects, ranging from methodological research on measurement and analysis of data to studies of the correlates of achievement. Goulet, et al. (1975) used the Anchor Test Study data to examine the severity of problems encountered in measuring achievement change, the feasibility of vertical test equating, and the stability of measurement constructs over time, in an extensive study of methodological problems in longitudinal reseach, supported by the National Institute of Education. Because some test forms used in the Anchor Test Study were recommended by their publishers for use with students in more than one of grades four, five and six, it was possible to examine the consistency of vertically equating relationships between levels of other tests, recommended for use in only one of these grades.

Data collected in the Anchor Test Study included a number of descriptors of the

schools and classrooms of students participating in the study, as well as descriptors of individual students. Principals provided information on the socio-economic composition of the attendance areas served by their schools and on the degree of urbanism of their schools' attendance areas. Teachers provided information on class size and the use of ability grouping, as well as descriptive information on individual students who participated in the study, such as IQ range, race, and primary language used in the student's home. Two studies (Burgdorf, 1976; Doucette and St. Pierre, 1977) made use of these data to examine some of the correlates of reading achievement in the upper primary grades. Burgdorf used Total Reading scores on the Metropolitan Achievement Test level administered to more than 65,000 fifth-graders to construct an extensive series of cross tabulations. He examined as many as three of the ten descriptive variables used in the Anchor Test Study in relation to distributions of Metropolitan Total Reading scores. In a similar study supported by the National Center for Educational Statistics, Doucette and St. Pierre (1977) found that school variables such as urbanism of school location, type of school support (public vs. private), socioeconomic composition of the student body, and percentage of minority enrollment were clearly related to reading achievement as measured by the Metropolitan Achievement Test. Likewise, individual student variables such as reported IQ range, race or ethnicity, primary language spoken in the student's home, and teacher's diagnosis of the existence of a reading problem were significant correlates of reading achievement. However, the two classroom variables studied, class size and presence or absence of ability grouping, were not found to be related to reading performance on the Metropolitan Achievement Test.

Rasp and Stiles (1976) and Rasp (1976) reported the results of a two-year experience in using the Anchor Test Study equating tables in conjunction with the Washington State Assessment Program. Instead of requiring the administration of a common reading achievement test throughout the state, the State Department of Education attempted to

develop a profile of reading performance for its fiscal year 1974 ESEA Title III needs assessment plan by analyzing sixth-grade achievement data routinely collected by a 20 percent sample of school systems. As might be expected, not all school systems used the test batteries, levels and forms equated in the Anchor Test Study, and problems of sampling bias were encountered. With National Institute of Education support, Rasp conducted a later study in which he developed computer programs that would apply the Anchor Test Study norm tables to data collected in statewide assessments. Experience gained in using Anchor Test Study results to aggregate statewide reading achievement data was applied to the development of generalizable guidelines for such applications.

In a National Institute of Education-supported study completed in 1980, Linn, et al. used Anchor Test Study data to investigate the possibility that content and format characteristics of reading comprehension items were consistently related to differences in item characteristic functions for students classified by race. This study of bias in reading comprehension items employed eight subgroups of students classified by grade level (fifth and sixth), income level of school attendance area (low and other), and race (black and white). A normative basis for judging meaningful differences in item characteristic curves was established by observing functional differences for groups of the same race that differed in grade level and income level of school attendance area. Unfortunately, the items that were identified as racially biased were not consistently different from other items in either format or content. Nonetheless, the Anchor Test Study provided a large and appropriate data base that enabled the authors to characterize item bias in a unique way and to examine substantive correlates of item bias.

Use in Equating Research. A considerable amount of research on the methodology of test equating has been completed using data from the Anchor Test Study. Because the Anchor Test Study data tapes contain extremely detailed information (in addition to demographic information, the raw data tapes identify the option chosen by every student

in response to every test item attempted), item characteristic curve equating models as well as methods that employ total test scores can be applied to the data set.

Rentz and Bashaw (1973; 1977) conducted an extensive reanalysis of all data collected in the restandardization and equating portions of the Anchor Test Study, using the Rasch one-parameter item response model. They estimated the Rasch difficulty of each reading comprehension and vocabulary item in the seven tests used in the original Anchor Test Study, and established common reference scales for vocabulary subtests and for reading comprehension subtests. Once the tests were placed on a common reference scale, Rentz and Bashaw determined corresponding raw scores on the vocabulary subtests of different test batteries, and corresponding raw scores on the reading comprehension subtests of different test batteries.

The Rentz and Bashaw study has important methodological implications for subsequent equatings of standardized achievement tests. Because the Rasch model purportedly provides "sample free" item calibrations, the representativeness of examinee samples used in the development of equating functions should not be a critical concern, as it is when classical equating procedures are used. It is also possible that sample size requirements will be somewhat smaller, since a more explicit relational model is being used.

Unfortunately, the results of the Rentz and Bashaw study were equivocal. For some of the subtests equated in the Anchor Test Study, the classical equating functions and the Rasch-determined equating functions were nearly identical over most of their score scales. For other subtest pairs, the differences between classical and Rasch equating functions were three or more raw-score points, an amount that is substantial when viewed as a component of bias error that will not diminish as a function of sample size. Even more perplexing is the question of which results are "correct" or "true." If one adopts the classical definition of equivalent scores -- scores that correspond to the same mid-percentile rank in any sample of examinees -- the classical equating function must be viewed as a standard. Conversely, if one defines as equivalent two scores that correspond

to the same Rasch ability level, the Rasch results must be viewed as a standard. For the moment, the classical definition of equivalent scores is more widely accepted.

Slinde and Linn (1977; 1978) used Anchor Test Study data to examine the feasibility of constructing consistent vertical equating tables for corresponding subtests in different levels of a test battery. They also examined the utility of the Rasch model in constructing vertical equating tables. They concluded that vertical equating of reading achievement tests was hazardous, regardless of the analytic method employed, and that use of different test levels in studies of achievement gain should be avoided if possible. Their findings also have implications for out-of-level testing, a practice that is common in evalations of compensatory education programs. Aggregation of achievement test data across levels of tests can lead to the incorporation of sizeable measurement errors.

With the support of the U. S. Office of Education, Bianchini and Vale (1975) examined the applicability of Anchor Test Study equating tables to groups composed of black or Spanish-surnamed students. They searched for evidence of interactions between equating relationships and the racial/ethnic composition of subgroups upon which they were based. Fortunately, no systematic relationships were found, and isolated evidence of an equating function by race/ethnic group interaction was attributed to relatively small black and Spanish-surnamed representation in the Anchor Test Study sample (leading to larger random equating errors), rather than to consistent racial/ethnic bias error.

Beard and Pettie (1979) used the Rentz and Bashaw (1973) reanalysis of Anchor Test Study data as a benchmark in judging the degree of Rasch fit of items in the Florida Educational Assessment tests in communications and mathematics administered to third-graders and fifth-graders. Their primary reaserch focus was a comparision of the results of classical linear equating and Rasch equating of test forms used in 1976 and 1977. They concluded that items in the Florida Asssessment tests fit the Rasch model to a greater extent than did items in the standardized achievment tests used in the Anchor Test Study.

It is interesting to note that they also found close correspondence between the results of linear equating and Rasch equating, suggesting (logically) that model fit may be critical when equating tests using the Rasch model.

Implications of the Anchor Test Study for Follow Through Research

From the review of literature reported above, it is clear that Anchor Test Study results have been used very little in large-scale program evaluation. Despite the supposition that the Anchor Test Study would facilitate collection of achievement test data that could be aggregated across projects, school systems, and states so as to provide bases for examining the targeting of services and the impact of services supported under Title I, ESEA, a thorough search of the ERIC system did not produce any supporting evidence.

Perhaps it is not surprising that the Anchor Test Study has contributed so little to state and federal evaluation of Title I, ESEA in view of the virtual abandonment of the large-scale survey approach to Title I evaluation at state and federal levels. At the time the Anchor Test Study was conceived, the U. S. Office of Education collected uniform data on the structure, organization, and operation of hundreds of Title I projects, as well as uniform information on the backgrond, characteristics, and participation of thousands of students. More than a few states emulated the federal approach to Title I evaluation. More recently, the Title I Evaluation and Reporting System has emphasized provision of data by local school systems using a common format, but allowing the use of any basic skills achievement measures that can be related to tests that have national norms. In effect, school systems have been encouraged to use criterion-referenced measures, and to equate these measures to nationally standardized tests through loosely controlled local equating studies. Much of the rhetoric of the measurement and evaluation community has served to relegate standardized tests to second-class status as instruments for use in program evaluation. Truly representative norms for the reading comprehension

and vocabulary subtests of the Metropolitan Achievement Tests, and tables of score correspondence between these subtests of the Metropolitan and those of seven other test batteries understandably hold less currency than they once did in the minds of evaluators.

As noted above, the research uses of Anchor Test Study data have far outpaced use of the study by evaluators. It is not unreasonable to conclude that the Anchor Test Study has led to a resurgence of interest in reasearch on test equating. Certainly, the research literature on test equating has grown at a far faster rate since the Anchor Test Study was completed than in the eight year period prior to publication of its final report. And a good bit of the empirical research on test equating completed in the last eight years has made use of the Anchor Test Study data tapes.

We have also noted the extensive use of Anchor Test Study equating tables and data tapes in secondary analyses and applications ranging from studies of the correlates of reading achievement to investigations of test item bias. The sheer size of the data base, in addition its nationally representative structure, has permitted the creation of large subsamples of examinees, classified on such variables as sex, race, IQ-level, and language usage. Such subsamples are essential to much empirical research on item bias and the correlates of achievement, thus supporting the conclusion that the Anchor Test Study greatly facilitated this research.

In view of the history of usage of Anchor Test Study results and data, it is reasonable to ask whether a similar study would be of material value either in the evaluation of the Follow Through program or as a part of the research in support of future Follow Through approaches envisioned by Schiller, et al. (1980). Each of these questions will be addressed separately. How could the results of a large-scale test equating study be used in Follow Through evaluation? Speculation on the usefulness of a test equating study in Follow Through evaluation must begin with the presumption that standardized achievement test results will, once again, constitute a primary indicator of

program effects. It should be clear that this presumption does not constitute a recommendation.

If standardized achievement tests ae not used to assess program impact, they might still provide a useful indicator of the characteristics of recipients of program services, or of the population of potential recipients. But this limited use of standardized achievement tests in Follow Through evaluation probably would not warrant the same attention to equivalence of measures as would use of such tests for assessment impact.

If a number of tests of basic skills suitable for use with children in kindergarten through grade three were to be equated successfully, the obvious advantage would be greater flexibility in the selection of tests for evaluation of basic skills programs at those grade levels. Potential benefits include savings in time and money, and increased measurement validity.

As noted above, two of the five generic program evaluation questions identified by Boruch and Cordray (1980), "Who is served by the program?" and "Who needs services?" could be answered in part through standardized achievement test reslts. Just as it is common to describe recipients and potential recipients of program services in terms of age distribution, racial and ethnic background, socioeconomic level, sex, urbanism of school, and grade level,,achievement status in the basic skills areas is another common descriptor. To secure such information, a standardized achievement test battery is typically administered to all program participants and, perhaps, to all students in selected grades in school systems that participate in a program. Often these same school systems administer one of a small number of standardized achievement test batteries as a part of their routine testing programs. As a result, students in the target grades of the compensatory program are subjected to two testing sessions, with consequent loss of instructional time, and demonstrably reduced motivation to perform well on the tests.

Were subtests in reading and mathematics from a number of widely used achievement test batteries to be equated successfully at levels suitable for use in kindergarten through grade three, future Follow Through evaluations might avoid special administrations of such tests for purposes of describing program participants, students in comparison groups, and potential program participants. Achievement test data already available in the archives of participating school systems could be translated to a common reference scale, and aggregated to form achievement distributions for all groups of interest.

The other obvious application of standardized achievement test data in Follow Through evaluation is in response to Boruch and Cordray's question "What are the effects of services on recipients?". Again, it is possible that equating of basic skills subtests at levels appropriate for use in kindergarten through grade three would allow the use of achievement test data already available in school systems' files to answer this question, with an attendant savings in testing time and testing cost. Since most Follow Through eligible school systems probably receive funds through Title I, ESEA, it is not unreasonable to expect that they routinely test some, if not all, of their students in the early elementary grades at the beginning and end of each school year.

Perhaps the greatests benefit of an equating of early elementary basic skills tests would be the possibility that Follow Through model sponsors could select one of a number of achievement tests for use in evaluating their models. The potential for increasing measurement validity is real and important, as evidenced by an increasng body of research on the effects of congruence between curriculum content, instructional content, and achievement tests. Although a variety of subtests carry the label "reading compehension test" or "arithmetic concepts test", it has become increasingly clear in recent years that they do not all measure the same thing. Bianchini (1978) addresses this problem in a paper on the appropriateness of differentiated norms for the evaluation of programs for disadvantaged and minority students. He cites the results of an analysis by Corder (1970)

of the finding that 65 percent of first-grade students in California scored below the first quartile on the national norms of the Stanford Reading Test when that test was used in an evaluation of the Miller-Unruh Reading Program in 1966. The California state legislature had budgeted the compensatory reading program on the basis of the reasonable expectation that about one-fourth of California's first-graders would score below the first quartile of the national norm distribution, and the disproportionate finding caused considerable reaction. In the midst of a variety of hypotheses on the reasons for the poor showing by California's first-graders --ranging from an analysis of the IQ distribution and racial composition of the sample used in the Stanford norms, to speculation about the excessive difficulty of Stanford Reading Test items -- Corder conducted an analysis of the congruence of the vocabulary assessed by the test and the vocabulary used in the state-provided instructional materials for first-graders. He found that the overlap was only 19 percent. On the basis of this finding alone, one cannot claim a causal realtionship. However, Bianchini completed a subsequent analysis of the vocabulary of first-grade readers used in California in 1971 and the vocabulary assessed by the Reading Test of the Cooperative Primary Test adopted by the state in that year for evaluation of the Miller-Unruh Reading Program. He found an overlap of 55 percent. Correspondingly, the performance of California's first-graders essentialy matched that of the national norms sample on the Cooperative Primary Tests. The medians matched perfectly, whereas the California median was at the thirty-eighth percentile of the Stanford Reading Test norms in 1966. Bianchini concludes:

> "The point is that in any program at the early grades it is particularly important for all children that the test content be related to instructional content. The reason for this is that children within the early grades learn only within the bounds of the curriculum they experience."

Further evidence on the need to consider the congruence between curriculum content

and test content in program evaluations was provided by Hoepfner (1978). After developing a taxonomy of content categories for reading tests and mathematics tests, Hoepfner categorized all items in the reading and mathematics subtests of the eight standardized achievement test series that were most widely used in 1978. He found that the subtests differed widely in their content emphases, despite their common titles. For example, at the levels recommended by their publishers for use with first-grade students, the percentage of items assessing mastery of word attack skills varied from a low of zero to a high of 60 percent in the reading subtests Hoepfner reviewed. In assessing recognition of word meanings (termed "vocabulary" in some tests), the percentage of items at the first-grade level varied from zero to 59 percent. And assessment of reading comprehension at the first-grade level was the function of 14 percent of the items in one subtest, of 53 percent of the items in another, and of widely varying percentages between these extremes in the remaining six.

Hoepfner found similar content differences in his review of mathematics subtests. For example, knowledge of numbers and sets was assessed by only two percent of the items contained in the second-grade level of one test and by 24 percent of the items contained in another test intended for second-graders. Whole-number computation was the objective of 60 percent of the items in one test intended for second-graders, but was not assessed at all by the items in the second-grade mathematics test of another battery.

In an analysis that was similar to Hoepfner's but involved multiple judges in the classification of items, Porter, Schmidt, Floden and Freeman (1978) found that the distributions of items in the mathematics subtests of standardized achievement batteries intended for use with fourth-graders differed substantially across objectives and content. For example, they found that operations with single-digit numbers were represented in only two percent of the items in one battery, but in 20 percent of the items in another. Problems involving whole numbers constituted 39 percent of the items in one battery, but made up 66 percent of the items in another. Addition varied from 12

percent of the items in one battery to 21 percent of the items in another. Graphs, figures, and tables were used as stimuli in 43 percent of the items in one battery, but were present in only 15 percent of the items in aother. Clearly, these tests differed substantially in their mode of presentation of arithmetic material, the arithmetic operations they required students to perform, and in the naure of the material they presented. Although their total-score intercorrelations would probably be in the high eighties, the tests would not provide equally valid representations of the effectiveness of a given basic-skills mathematics program.

Porter, et al. conclude as follows (p.538):

> "Treating practical significance in instructional program evaluation requires intimate familiarity with the measures on which effects are estimated and their substantive relationship with the goals of the program being evaluated. Past attempts to provide general solutions to the size of effect problem have relied on standardized indices which can be estimated and reported without any knowledge of what was measured. For this reason these efforts are viewed as steps in the wrong direction. Instead, what is called for is a procedure whereby the substantive goals of the program, the instructional outcomes implied by a test, and the interrelationship between the two are made explicit. The procedure should facilitate investigation of treatment-by-item interactions and at the same time facilitate a description of the measures in sufficient detail to support inferences regarding practical significance."

As has been discussed earlier, previous national evaluations of the Follow Through program have fallen prey to the error that Porter and his colleagues identify. If a number of widely-used standardized achievement batteries suitable for use in kindergarten through grade three could be equated, proogram sponsors could then select reading subtests and mathematics subtests from any of the equated batteries for use in evaluating

their Follow Through models. Even with the diversity of content emphases noted above, it is not certain that the curriculum congruence of any of the subtests would be adequate to the evaluation of all (or even most) Follow Through models. However, it is far more likey that suitable content matches between tests and curricula could be realized if seven or eight test batteries were available, than if one one battery was used for evaluation of the entire program, as was the case in the SRI-Abt evaluation. This advantage alone might justify NIE's investment in another large-scale test equating study.

How could the results of a large-scale test equating study be used in Follow Through Research?

The program of inquiry on early primary education for children from low income families envisioned by Schiller, et al. (1980) includes the desire to develop:

"New uses for information systems, including testing and evaluation results, to bring better diagnostic and prescriptive information to bear on Follow Through student learning needs." (p.11).

In a variety of other sections, the planning document recognizes the need to develop new strategies and procedures for assessing the consequences of Follow Through interventions.

A large-scale test equating study has the potential of contributing to a better understanding of the outcomes and effects of early primary education programs in a number of ways. Some of the research outcomes might provide tools that could be applied directly to the future evaluation of Follow Through and other early primary intervention programs; while other research products would be more fundamental and less immediately applicable. A good bit of the research would focus on test equating methodology itself; while other foci would involve extensions of the research that has eminated from the Anchor Test Study.

It is not clear that widely used standardized achievement tests appropriate for students in kindergarten through grade three are similar enough in their psychometric

characteristics to allow successful equating. An initial benefit of a test equating study at these grade levels would be an examination of the feasibility of equating various early primary reading tests and various early primary arithmetic tests. Such a feasibility analysis would extend current knowledge on the degree of parallelism required to sustain consistent equating relationships between non-parallel tests.

The literature on test equating (Lord, 1950; Angoff, 1971; Jaeger, 1981) suggests that, although non-parallel tests can be calibrated, they cannot be equated. The distinction is in the consistency of the scaling relationship between the two tests across various populations of examinees. Strictly parallel tests are virtually interchangeable, in that they measure the same psychometric function with the same degree of reliability for all groups of examinees. For strictly parallel tests, then, an equating relationship is unique. Once established for any population of examinees, it holds for all populations. In contrast, non-parallel tests differ either in the psychometric function measured, in reliability, or in both characteristics. Although it is possible to establish a functional correspondence between the scales of non-parallel tests (e.g., by defining as equivalent, raw scores corresponding to the same standard score or the same percentile rank) for a given population, the relationship will not be consistent for another population. In theory, then, the original Anchor Test Study should have been infeasible.

In practice, the alternate forms of standardized achievement tests produced by virtually all test publishers are not strictly parallel. Although they constitute the best approximations to parallel forms presently available, they differ to some degree in overall difficulty, raw-score variability, and internal consistency reliability. The intercorrelations of alternate forms are extremely high, but are often slightly lower than their internal consistency reliabilities would allow. Thus strict parallelism is a theoretical ideal that is approached, but never realized in practice. Yet alternate forms

of standardized achievement tests are routinely equated, and the equating relationships established appear to be acceptably consistent.

If properly designed, a large-scale test equating study at the early primary grades would support an analysis of the content similarity requirements of non-parallel tests and correlational requirements of non-parallel tests in order to achieve equating realtionships that were sufficiently consistent over populations that differed in socioeconomic composition, IQ distribution, racial composition, and other demographic descriptors that typically distinguish low income students from the majority of students, that they could be used in large-scale evaluation studies or for individual assessments. Bianchini and Vale (1975) have completed an initial exploration of the parallelism requirements of successful equating using data from the Anchor Test Study. But their findings are limited in two ways. First, they apply only to reading comprehension and vocabulary subtests appropriate for use in grades four through six. Second, the sampling procedures used in the Anchor Test Study sought proportional representtation of students in various minority ethnic and racial groups, and therefore sampled such students in far smaller numbers than majority white students. Differences in equating relationships found for white students and black students could as readily be attributed to random fluctuations as to consistent bias errors. Nonetheless, Bianchini and Vale concluded that the equating relationships developed in the Anchor Test Study were reasonably consistent across racial groups, and recommended that the equating tables be used with black students and Spanish-surnamed students, as well as with white students.

The results of an additional equating study could be used to test the limits of generalizability of this finding across grade levels and across subject areas. In addition, a new test equating study cold be designed so as to sample racial and ethnic minorities, and students in other groups of interest, in sufficient numbers to allow clear differentiation between random fluctuations, between equating relationships and truly stable inconsistencies.

One other important consequence of the intercorrelation between tests to be equated is the stability (degree of random fluctuation across samples from the same population) of equating relationships. An explicit mathematical relationship between inter-test correlation and the standard error of equating has been developed for classical linear equating (Lord,1951). However, similar relationships are not available for the form of equipercentile equating found to be most stable in the Anchor Test Study, nor for equating procedures that employ item response theory models. An additional test equating study would have the potential of greatly extending the empirical basis for examining the relationship between test characteristics aand the statistical stabilty of equating functions.

As was the case for tests equated in the Anchor Test Study, publishers of the most widely used standardized achievement tests recommended for use in the early primary grades suggest various combinations of levels of their tests for students in the gradespan of interest (Hoepfner, 1978). For example, the same level of the California Achievement Tests is recommended for use in grades one and two, and a different level is recommended for use in grade three. However, three different levels of the Iowa Tests of Basic Skills are recommended for use in grades one, two and three. If publisher's recommendations are followed in the administration of various test levels in the early primary grades, an equating study at those grade levels would provide the data necessary to conduct research on the consistency and feasibility of vertical test equating across levels of standardized tests. As noted above, Slinde and Linn (1977; 1978) have examined vertical equating relationships for reading comprehension tests at grades four through six. The methodology they have developed for this type of research could be applied directly to tests suitable for use in another gradespan and to tests in another subject area.

The relative utility of classical equating models vs methods that employ various

item response theory models is subject to debate, despite extensive research on the topic (Beard and Pettie, 1977; Marco, Petersen and Stewart, 1980). Although Beard and Pettie concluded that: "Rasch equating was comparable to linear equating in the analysis of longitudinal trends in basic skills achievement."; Marco, et al. found substantial differences between the methods in terms of bias error and random error, depending on the comparability of the tests being equated and the comparabilty of the samples of examinees used to gather data for equating. Beard and Pettie employed items from the Florida Assessment tests in commmunications and mathematics, appropriate for third-graders and fifth-graders, whereas Marco, et al. based their analyses on Scholastic Aptitude Test items ad samples of high scool students. Differences in their conclusions are likely attributable, at least in part, to their use of items from different tests and examinees in different gradespans. It appears that various equating methods will produce similar results under some circumstances and substantially different results under others. The number of variables involved in the relationships among various equating methods is such that purely analytic rules of correspondence are not likely to be developed soon. An intriguing question that is yet to be resolved, as noted above, is the choice of an appropriate standard for judging the correctness of an equating relationship. That is, if two equating methods produce substantially different results, which one should be regarded as correct? Although this logical definitional problem is unlikely to be resolved through another large-scale equating study, the circumstances in which various equating methods yield emparable results or divergent results could be further explored using data from an equating study involving basic skills tests with early primary students. The methodology that Marco, et al. have applied to the Scholastic Aptitude Test data base could be employed directly with early primary equating results and, in parallel fashion with results from the original Anchor Test Study. The major outcomes of this research would be greater understanding of the conditions needed to equate two tests successfully; conditions under which one equating method might produce

stable and consistent equating functions, whereas aother might not; and conditions under which various equating methods, including classical and item response theory methods, produce virtually identical results.

Beyond its potential value in fostering additional research on test equating, another major test equating study could provide a data base that would facilitate extension of the more general research that has emerged from the original Anchor Test Study. In particular, the research cited above that concerns the correlates of academic achievement, an area of major research emphasis in the 1981 NIE Research Grants Announcement on Testing and Evaluation, could be extended to the early primary grades. Although some interesting findings on the correlates of achievement emerged from the Anchor Test Study, they were limited by the restricted range of ancillary data collected in that study. Since there was no intention of using Anchor Test Study results in an investigation of the correlates of ahievement at the time the study was designed, the only ancillary data collected were those needed to verify the representativeness of samples used or to examine the comparability of equating relationships across sex, race, and IQ groups. A new equating study could be carefully designed to support a far greater range of investigations, including studies of the correlates of achievement, and would therefore be of greater value than the Anchor Test Study in terms of secondary analyses.

As noted above, another area of research that has made use of Anchor Test Study results is analysis of test item bias. Again, a test equating study at the early primary grades could be designed so as to facilitate this objective. Careful attention to the adequacy of samples of students of various minority groups and IQ groups, and students of both sexes would be required, as is the case in many of the research areas already discussed. In addition, data tapes would have to be designed to allow recovery of the most basic information on students' responses to test items, as was done in the Anchor Test Study.

In summary, a new test-equating study at the early elementary grades would be consistent with several Follow Through objectives. First, it would have the potential of contributing to the validity of an evaluation of Follow Through effects (assuming that the past and present policy of using standardized achievement test results as a primary indicator of Follow Through success is continued). Second, such a study would contribute to research on methods of assessing programs for children of low-income families at the early primary grades through fundamental research on test equating methods and through research in a variety of ancillary areas.

RECOMMENDATIONS

Although the potential benefits of a test equating study at the early primary grades have been discussed in some detail, it is not recommended that such a study be initiated as a part of Follow Through research without additional planning and investigation. In particular, use of standardized achievement tests in the early primary grades may well have diminished considerably since Hoepfner's study was completed in 1971 (the year of the USOE-sponsored conference at which he presented the paper published in the 1978 reference). If so, the value of an equating study at the early primary grade levels would be reduced, apart from its utility in large-scale program evaluations that incorporated standardized achievement tests. Further, patterns of test usage may have changed since Hoepfner identified the eight test batteries used in his study as those most widely used. Reliable information on current test usage would be needed prior to selection of test batteries for an equating study, and to provide a basis for deciding whether or not a test-equating study at the early primary grade levels was warranted.

If a review of the use of standardized achievement tests in the early primary grades suggested that relatively few tests were widely used in the basic skills areas, a study of the feasibility of equating corresponding subtests of those test batteries would be a logical step. Administration of pairs of subtests to samples of examinees large enough to

estimate inter-test correlations and internal consistency reliabilities would provide the information needed to make a reasoned decision on whether to conduct a large-scale equating study.

REFERENCES

Anderson, R.B., St.Pierre, R. G., Proper, E.C., & Stebbins, L.B. Pardon us, but what was the question again? Harvard Educational Review, XLVIII, May, 1978, 161-170.

Angoff, W.H. Scales, norms and equivalent scores. In R.W. Thorndike (Ed.) Educational Measurement (2nd. ed.), Washington, D.C.: American Council on Education, 1971, 508-600.

Armbruster, B.B., Stevens, R.J., & Rosenshine, B. Analyzing content coverage and emphasis: a study of three curricula and two tests. Technical Report No. 26. Urbana, IL: Center for the Study of Reading, University of Illinois, March, 1977, 22pp.

Beard, J.G. & Pettie, A.L. A comparison of linear and Rasch equating results for basic skills assessment tests. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, April, 1979, 23pp.

Bianchini, J.C. Achievement tests and differential norms. In Wargo, M.J. & Green, D.R. (Eds.) Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Monterey, CA: CTB/McGraw-Hill, 1978, 157-181.

Boruch, R.F. & Cordray, D.S. An appriasal of educational program evaluations: federal, state and local agencies. Evanston, IL: Northwestern University, June, 1980.

Burgdorf, K. Anchor test study. Fifth-grade achievement as a function of selected school, classroom, and pupil variables. Rockville, MD: Westat Research,Inc. 1976, 100pp.

Buros, O.K. Fifty years in testing: some reminiscences, criticisms and suggestions. Educational Researcher, 1977, 6, (7), 9-15.

Chang, S.S. & Raths, J. The school's contribution to the cumulating deficit. Journal of Educational Research, 1971, 64, 272-276.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A., Weinfeld, F.D., & York, R.L. Equality of educational opportunity. Washington, D.C.: U.S. Government Printing Office, 1966.

Cooley, W.W. & Leinhardt, G. The instructional dimensions study: the search for effective classroom processes. Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh, August, 1978, 91pp.

Corder, R. The Cooperative Primary Reading Test and its relationship to the California Miller-Unruh Program. Berkeley, CA: Unpublished research memorandum. Educational Testing Service, 1970.

Doucette, J. & St.Pierre, R. Anchor test study: school, classroom, and pupil correlates of fifth-grade reading achievement. Cambridge, MA: Abt Associates, 1977, 232pp.

Goulet, L.R., et al. Investigation of methodological problems in educational research: longitudinal methodology. Final Report. Urbana, IL: University of Illinois, 1975, 261pp.

Hoepfner, R. Achievement test selection for program evalation. In Wargo, M.J. & Green, D.R. (Eds.) Achievement Testing of Disadvantaged and Minority Students for Educational

Program Evaluation. Monterey, CA: CTB/McGraw-Hill, 1978,277-324.

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

Jaeger, R.M. The national test-equating study in reading. Measurement in Education, 4. (4), summer, 1973

Jaeger, R.M. Comments on Coffman's paper. In Wargo, M.J. & Green, D.R. (Eds.) Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation. Monterey,CA: CTB/McGraw-Hill, 1978, 120-128.

Jaeger, R.M. The effect of test selection on Title I project impact. Educational Evaluation and Policy Analysis, 1, (2), Spring, 1979

Jaeger, R.M. Some exploratory indices for selection of a test-equating method, Journal of Educational Measurement, in press, 1981.

Jenkins, J.R. & Pany, D. Curriculum biases in reading achievement tests. Technical Report No. 16. Urbana,IL: Center for the Study of Reading, University of Illinois, November, 1976, 24pp.

Linn,R.L. The anchor test study: the long and the sort of it. Journal of Educational Measurement, 1975, 12, 201-214

Linn, R.L., et al. An investigation of item bias in a test of reading comprehension. Technical Report No. 163. Cambridge, MA: Bolt, Beranek and Newman, Inc., March, 1980, 97pp.

Lord, F. M. Notes on comparable scales for test scores. Research Bulletin No. 43, Princeton, NJ: Educational Testing Service,1950.

Loret, P.G., Seder, A., Bianchini, J.C., & Vale, C.A. Anchor test study: final report. Berkeley, CA: Educational Testing Service, 1972

Loret, P.G., Seder, A., Biachini, J.C., & Vale, C. A. Anchor test study supplement: final report. Berkeley, CA: Educational Testing Service, 1973.

Porter, A.C., Schmidt, W.H., Floden, R.E. & Freeman, D.J. Practical significance in program evaluation. American Educational Research Journal, 15, (4), 1978, 529-539.

Rasp, A. Jr. & Stiles, R. Using the anchor test study in state assessment. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, April 19-23, 1976, 9pp.

Rasp, A. Jr. Using anchor test study tables in state assessment programs. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement and Evaluation, ERIC-TM-58, 1976,8pp.

Rentz,R.R. & Bashaw, W.L. Equating reading tests with the Rasch model. Athens, GA: Educational Research Laboratory, University of Georgia, August, 1973.

Rentz,R.R. & Bashaw, W.L. The national reference scale for reading: an application of the Rasch model. Journal of Educational Measurement, 14, (2),1977, 161-79.

Schiller, J., Stalfort, C., Rudner, L., Kocher, C. & Lesnick, H. Plans for Follow Through

research and development, Washington, D.C.: National Institute of Education, October, 1980.

Schutes, R.E. Verbal behaviors and instructional effectiveness. Unpublished Doctoral Dissertation, Stanford University, 1969.

Slinde, J.A. & Linn, R.L. Vertically equated tests: fact or phantom? Journal of Educational Measurement, 14, (1), 1977, 23-32.

Slinde,J.A. & Linn, R.L. An exploration of the adequacy of the Rasch model for the problem of vertical equating, Journal of Educational Measurement, 15, 1978, 21-35.

Stebbins, L.B., St.Pierre, R.G., Proper, E.C., Anderson, R.B. & Cerva, T.R. Education as experimentation: a planned variation model. Volume IV-A, An evaluation of Follow Through. Cambridge, MA: Abt Associates, Inc. 1977.

Stonehill, R.M. and Fishbein, R.L. Summarizing the results of Title I evaluations -- the comparability of achievement gains. Paper presented at the Conference on Large-Scale Assessment, Denver, CO, June 11-14, 1979,14pp.

Tittle,C.K. Judgmental methods in test development. Presented at the 1980 Johns Hopkins University National Symposium on Educational Research: Test Item Bias Methodology, the State of the Art: Washington, D.C., November 7, 1980.

Tyler, R.W. Council of Basic Education Bulletin, March, 1972

Tyler, R.W. Comments on Hoepfner's paper. In M.J. Wargo and D.R. Green (Eds.) Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation. Monterey, CA: CTB/McGraw-Hill, 1978, 324-329.

Vale, C.A. & Bianchini, J.C. Evaluation of educational achievement test measures as an eligibility criterion in the Better Schools Act formula. Final Report. Berkeley, CA: Educational Testing Service, 1978, 64 pp.

Wisler, C.E., Burns, G.P.Jr., & Iwamoto,D. Follow Through redux: a response to the critique by House, Glass, McLean and Walker. Harvard Educational Review, XLVIII, May, 1978, 171-185.