DOCUMENT RESUME

ED 145 011 UD 017 296

AUTHOR McLaughlin, Donald H.; And Others
TITLE Controversies in the Evaluation of Compensatory Education
INSTITUTION American Institutes for Research in the Behavioral Sciences, Palo Alto, Calif.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO AIR-61700-7-77-FR-II
PUB DATE Jul 77
NOTE 151p.; For related documents see UD 017 295 and UD 017 297

EDRS PRICE MF-$0.83 HC-$8.69 Plus Postage.
DESCRIPTORS *Compensatory Education; *Data Analysis; *Measurement; Measurement Goals; Measurement Techniques; *Methods Research; *Research Design; *Research Methodology; Research Problems; Sampling
IDENTIFIERS *Elementary Secondary Education Act Title I; ESEA Title I

ABSTRACT

This document answers the question, "What has been learned about evaluation methodology from the decade of compensatory education"? Ten issues dealing with the evaluation of Title I were identified within a general theoretical framework of evaluation. For each issue it was the aim of this document to do the following: 1) to clarify the issue, 2) to point out examples in which it is crucial, 3) to present and evaluate arguments on different sides of the issue, and 4) to suggest resolutions of the issue. Each of the issues was selected because its resolution is a necessary step in the development of a rational Title I evaluation policy. The issues addressed are: 1) To what should Title I treatments be compared? 2) Is longitudinal evaluation necessary? 3) When is representative sampling important? 4) How large a sample is necessary? 5) What constructs should be measured to determine Title I impact? 6) What types of achievement measurement instruments should be used in Title I evaluation? 7) What units of measurement should be used? 8) What are the conditions for valid comparisons between nonequivalent treatment and comparison groups? 9) Under what conditions can relationships of Title I costs and treatments to effectiveness be inferred? 10) How should data be aggregated across projects in Title I evaluations? (Author/AM)

AIR-61700-7/77-FR II

CONTROVERSIES IN THE EVALUATION OF COMPEN 'ATORY EDUCATION

Donald H. McLaughlin
Kevin J. Gilmartin
Robert J. Rossi

July 1977

AMERICAN INSTITUTES FOR RESEARCH
Post Office Box 1113 / Palo Alto, California 94302

CONTROVERSIES IN THE EVALUATION OF COMPENSATORY EDUCATION

Donald H. McLaughlin
Kevin J. Gilmartin
Robert J. Rossi

July 1977

This report is a product of contract #400-76-0129 between the National Institute of Education and the American Institutes for Research. Its content reflects the view of the individual authors solely.

Preface

In accordance with the objectives of contract #400-76-129 between the National Institute of Education and the American Institutes for Research, this document was produced to present as clearly as possible to non-methodologists the sources of controversy surrounding the decade of evaluation of Title I of the Elementary and Secondary Education Act of 1965. At the beginning of October 1976, the authors set out to write (1) this methodological discussion, (2) a volume of summaries of Title I evaluation studies, and (3) a synthesis of the substantive findings about Title I that have been provided by those studies. Allotment of time to the three documents was a constant problem during the nine-month period of the contract, because the direct expenses covered only about 10 person-months of professional effort. This document represents about 40% of that effort.

The authors intended to produce a document that would serve as the first draft of an introductory textbook for educational evaluation. While this goal would, we feel, fulfill a real need, it has proven more difficult than expected to explain complex problems simply, and we would welcome any readers' suggestions on how to do that better. The current document contains very little algebraic notation; however, lapses into undefined technical jargon can be frustrating to readers who are completely unfamiliar with the subject matter, and we cannot guarantee to have eliminated all such lapses.

The first draft manuscript for this document was produced in February 1977 to delimit the scope of the task. It was circulated among the authors and to the NIE project monitor, Alison Wolf, whose comments on this draft were very helpful. A second draft was produced in March 1977 and circulated to several reviewers. These reviewers were exceptional in their donation of time to this endeavor and the sophistication of the feedback they provided. They were Michael Wargo, now the Director of the Evaluation Division of the Office of Policy and Planning in ACTION: G. Kasten Tallmadge of RMC Research Corporation; Jane David of the Educational Policy Research Center at the Stanford Research Institute; Alison Wolf and Joy Frechtling of NIE; and from the staff of AIR, William Clemans, William Shanner, and Marion Shaycoft. We are deeply grateful to these individuals for their efforts; however, because we did not follow their counsel in every case, we accept responsibility for any faults that remain in this document.

The third draft of this document, essentially that which is presented here, was completed in early July 1977 and was sent to NIE for final approval prior to printing. The authors wish to express their special thanks to Alison Wolf for her understanding of the budgetary and temporal constraints involved in producing this document as well as for her cogent advice on improving the content and format of the document.

Finally, we are very grateful for the exceptional efforts of Ms. Emily Campbell, who gracefully accepted our missed interim deadlines and efficiently turned our manuscript into a presentable document through her typing/editing expertise.

Donald H. McLaughlin
Kevin J. Gilmartin
Robert J. Rossi

TABLE OF CONTENTS

Table of Contents, continued

Page

Summary (continued)

LIST OF TABLES AND FIGURES

CONTROVERSIES IN THE EVALUATION OF COMPENSATORY EDUCATION

Introduction

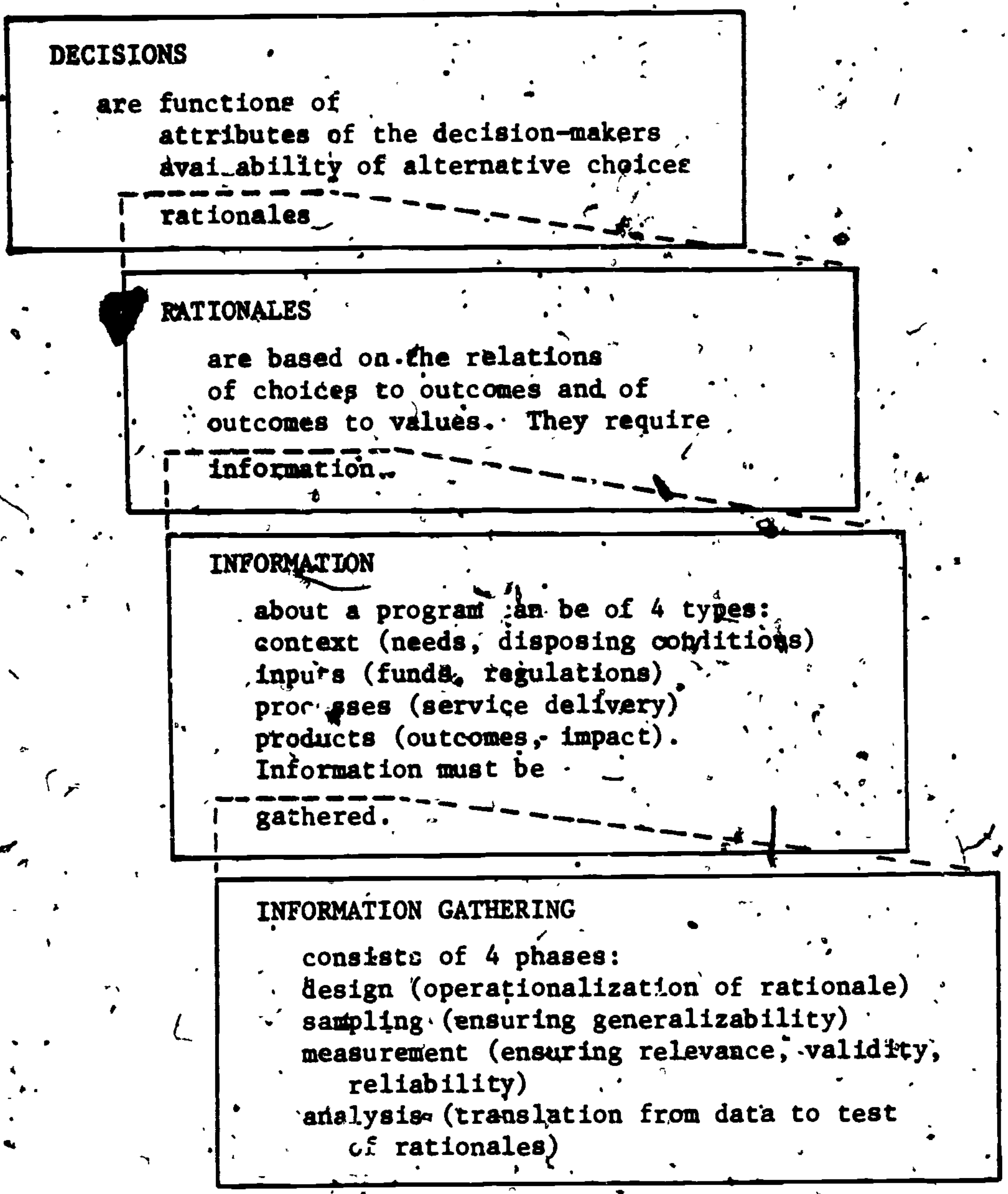
Since the middle 60s, many billions of dollars have been allocated through the federal government to social action programs, and many millions have been spent on the evaluation of these programs. In particular, the 15 billion dollars that have been spent on compensatory education through Title I of the Elementary and Secondary Education Act of 1965 have been accompanied by a continual stream of evaluation efforts. As has been pointed out by several authors, program evaluation is in a sense an adversary of program operation, and throughout the last decade there has been a great deal of criticism of programs by evaluations and also criticism of the evaluations by proponents of the programs. In this document, we would like to set forth the critical issues in the evaluation of compensatory education and attempt to supply the reader with an understanding of the complexities involved so that he or she can judge how and why to do evaluations as well as the validity of others' evaluations.

The crucial issue, as set forth in the classic argument between Donald Campbell and John Evans (Campbell and Erlebacher, 1970; Evans, 1970) is whether evaluations should be done perfectly or not at all, or should be done as well as possible in each situation. On the one hand, federal programs will inevitably be evaluated during congressional subcommittee presentations, whether based on quantitative data or on anecdotes, so it seems prudent to provide as much valid, objective, representative information as possible to our policy decision-makers. On the other hand, evaluation carried out by credentialed scientific organizations and academic institutions carries some weight thereby and correspondingly reflects on their reputations, and providing the stamp of scientific integrity to a compromised evaluation may result in the end in the debasing of the scientific method. If an evaluation must itself be evaluated before acceptance (Scriven, 1976), the resulting infinite regression ensures the lack of value of evaluation as a tool in policy-making. Evaluations must be carried out by proficient investigators with proper objectives, and the audience of the reports must be sufficiently aware of the issues in evaluation to judge for themselves that the evaluations are performed acceptably.

Ten issues dealing with the evaluation of Title I have been identified, within a general theoretical framework of evaluation. For each issue, it is the aim of this document (1) to clarify the issue, (2) to point out examples in which it is crucial, (3) to present and evaluate arguments on different sides of the issue, and (4) to suggest resolutions of the issue. Each of the issues was selected because its resolution is a necessary step in the development of a rational Title I evaluation policy.

This document discusses evaluation in the context of decision-making within a rational planning system. As shown in Figure 1, the system has four primary components: *decisions*, *rationales*, *information*, and *gathering* (of *information*). *Decisions* have *rationales*; and *information* is in turn *gathered* to test and validate these *rationales*. The term "evaluation" can refer to either the total system or subsets of it, although it is usually limited to the gathering of information. The decision-theoretic approach (Edwards, Guttentag, and Snapper, 1975) is the clearest example of the widest scope of evaluation in in this framework. From that viewpoint, the task of a program evaluation includes, among other things, the analysis of the decision process, and in particular, the quantitative determination of values that affect decisions. Strict adherence to this framework would exclude from consideration research studies whose product is not related to decisions (e.g., basic research to determine the nature of educational disadvantage) and studies called "evaluations" but undertaken for extraneous purposes (discussed by Floden and Weiner, 1976). However, that fact will not preclude the discussion of such studies as they relate to Title I in this document.

The separation of the four primary components of rational planning is an important step in the identification of different types of evaluations. Evaluations can be characterized by the types of decisions to be made, the types of rationales advanced for them, the types of information relevant to testing and validating the rationales, and the ways of gathering the information. The most notable distinction of evaluations in terms of decision type is the formative-summative dichotomy (Scriven, 1967); according to that dichotomy, information gathered in evaluation can be used either to improve the process evaluated (formative evaluation) or to support a decision of whether to make further investment in the process evaluated (summative evaluation). That distinction affects all components to the extent that the type of decision determines the rationales, the information needed, and the appropriateness of ways of gathering information.



DECISIONS

are functions of
attributes of the decision-makers
availability of alternative choices
rationales

RATIONALES

are based on the relations
of choices to outcomes and of
outcomes to values. They require
information.

INFORMATION

about a program can be of 4 types:
context (needs, disposing conditions)
inputs (funds, regulations)
processes (service delivery)
products (outcomes, impact).
Information must be
gathered.

INFORMATION GATHERING

consists of 4 phases:
design (operationalization of rationale)
sampling (ensuring generalizability)
measurement (ensuring relevance, validity, reliability)
analysis (translation from data to test of rationales)

Figure 1. Schematic diagram of the framework of evaluation

A plausible rationale for any decision must take the form of an argument that the value of the expected outcome given one choice is greater than the values of expected outcomes given other choices. Independently of whether the link drawn between a decision and later outcome is correct, there can be substantial disagreement about which aspects of outcomes are to be considered. A good deal of controversy over evaluation stems from this fact. The need for information gathering arises when a rationale contains an empirically testable statement whose truth is in question (e.g., statements like "if we can get the money translated into smaller student:teacher ratios, achievement gain will grow").

The four types of information shown in Figure 1 correspond to the four types of evaluation identified by Stufflebeam (1971) and referred to frequently as the CIPP model of evaluation. Information relevant to a particular decision rationale may pertain to a program's context (e.g., the needs and abilities of the target group), to its inputs (e.g., the funding pattern and regulations), to its processes (e.g., the selection of participants and of treatment methods and the implementation of treatments), and to its products, or outcomes. The products of a program to be evaluated can be expected to vary along a proximal-distal continuum: proximal outcomes tend to be more under the control of the program to affect and less subject to contextual factors, whereas distal outcomes tend to be more clearly related to values which programs are hoped to achieve. Stufflebeam pointed out the ways in which each of the four types of information is especially important for a particular decision type; of course, the four types of information are useful in combination for many decisions.

The four aspects of gathering information form the methodological substance of most evaluations of federal programs, as reported by the researchers who carried out the studies. The methodological issues to be discussed in the present document will be presented in four sections corresponding to these aspects of information gathering. Design issues refer to problems in the general plans for testing of decision rationales. In many actual cases, the rationales to be tested have not been made explicit and can only be inferred from the nature of the report's conclusions. Sampling issues refer to problems in generalizing to a population, and they concern the size, representativeness, and units of the sample. Measurement issues refer to problems in translating fundamental program concepts (e.g., "educational disadvantage") into instruments to assess the concepts and to problems in the assignment of numerical scores to recorded behaviors (scaling). Finally, analysis issues refer to problems in isolating and explaining particular relations in the data.

Before launching into the discussions of methodological issues, we shall provide some context by expanding the general evaluation framework of Figure 1 as it applies to federal studies of compensatory education. Each of the issues in the four areas will be discussed abstractly, as it pertains to any potential evaluations of Title I, and it will also be discussed in terms of specific past evaluations for which it is relevant. The inclusion of particular projects as examples in the discussions will take the projects out of context, however, and readers should not consider these discussions to constitute evaluations of the projects. Finding that a project has some methodological weakness may not diminish the importance of many of its conclusions, especially for studies that address many different aspects of Title I.

Decisions and Rationales in Title I

The primary decision-makers in the Title I system are Congress, the U.S. Office of Education (USOE), local school administrators, and teachers. Although each of these groups is far from monolithic and makes numerous and diverse decisions that affect the operation of Title I, it is helpful to lay out nine of the major decision types they address and the rationales and information needs for them: four decisions by Congress, three by USOE, and one each by local administrators and teachers.

1. Congress decides whether to increase the appropriation level. There are at least three basic rationales for increasing funds: (a) the program is reaching only some of the intended target population, it is helping those reached, and the reason it is not reaching others is because there are too few funds; (b) an effective method for solution has been found, but its typical per-pupil cost of implementation is higher than the typical expenditure allotted to each participant; or (c) increased costs for the same services require increased expenditures. Although the reason any particular member of Congress votes to increase Title I appropriations is a complex function of competing forces that may involve decisions on other appropriations completely unrelated to compensatory education, any decision to increase Title I funding must be accompanied by a rationale such as those listed--otherwise, it can be attacked as irrational or as an instance of "boondoggling."

   Even though we cannot hope to compile a complete set of rationales here, those that are included serve to identify the types of information needed. For the first rationale to be useful, Congress must know who the target population is and what the discrepancies are between the target population (educationally and economically disadvantaged children) and the participant population. They must also know whether Title I is helping those it reaches. For the second rationale, Congress must know of methods found to be effective and

and capable of being widely utilized, their costs, and typical per-pupil allocations. For the third rationale, Congress must know how inflation contributes to the costs of compensatory education services and what the effects would be of "holding the line on spending."*

A comprehensive evaluation of Title I would aim to provide Congress with the information necessary to test the validity of the various rationales. Due to constraints of time and effort, however, evaluations normally provide only partial validation of rationales, which, although it is useful, leaves significant gaps to be filled by faith. An example related to the first rationale above would be a study that demonstrated that substantial numbers of disadvantaged children were not being served, but failed to demonstrate that the children who were served benefited from the service. Whenever it is infeasible to close the informational gaps completely, an evaluation will be most useful when it addresses the gaps with whatever information is available.

In discussing this first decision type, we have tried to explain some of the problems that arise in relating decision-making to information-gathering. These apply also to the remaining nine decision types, although they will not be presented in equal detail.

2. <u>Congress decides whether to decrease the appropriation level</u>. The rationales for decreasing spending are not merely the inverses of the rationales for increasing spending. Two rationales for this decision might be (a) that funds are being used for services for people other than disadvantaged children or (b) that the need for a federal compensatory education program had diminished. Another possible rationale, that although the need persists the program is not dealing with it, is an argument for changing the program, not reducing its funding level. To test the two rationales for decreasing funding, the necessary information includes the distribution of compensatory education needs and services throughout the country.

*This third rationale is relatively weak, because it can be applied to all appropriations.

3. Congress might modify the funding allocation formula. The rationale for this decision might be either (a) that the children served by Title I are not exactly those for whom the program was intended or (b) that the nature of the need served by the program is modified. Again, the necessary information concerns the distribution of compensatory education needs and services, but possibly with emphasis on variations among needs and services.

4. Congress might modify or add a rule concerning the use of program funds. The rationale for this decision would be the identification of a problem that reduces the effectiveness of the program and a general method for eliminating or reducing the frequency of that problem. The information needed for this type of decision is therefore evidence that a particular unintended process frequently occurs in implementation of the program and that this process reduces program effectiveness. The latter type of evidence is necessary in order to avoid eliminating effective processes, and its validity depends upon the demonstration of causal linkages, not merely correlations: it is quite likely that the program will be more effective in some situations than in others but the situation is not the cause of the effectiveness. Another type of evidence, that a particular modification to the law will deal effectively with the problem, is unlikely to be available before the modification is made, but can be obtained after the modification by comparing the prevalence of the problem before and after the modification. Further modification can then be made.

Turning now to the U.S. Office of Education, we have three more major decision situations. One of these is essentially the same as the congressional decision to modify or add a rule.

5. USOE might modify or add a rule concerning the use of program funds. That rule might be in the form of a regulation (with the status of a legal requirement) or a guideline (a formal suggestion for procedures). The rationale and evidence necessary for such a decision would be the same as for the analogous congressional decision.

6. USOE may decide to disapprove a state's application for its annual allotment of funds or to request return of funds. The rationale

for this decision would be that the state is not complying with the law and regulations, that its noncompliance reduces the effectiveness of the program, and that punitive action would be likely to improve program performance. Evidence needed for this rationale concerns processes and outcomes within particular states and local districts, rather than national averages. It also concerns whether punitive action will deal with the problem, which, except for generalization from other federal programs, can only be determined after the action is taken.

7. USOE may decide to provide technical assistance. In fact, that decision may be incorporated into the law by Congress, as in the case of the instructions to USOE to provide technical assistance to states and local districts in the preparation of their annual Title I evaluation reports. The rationale for such a decision is that there is a clear and pervasive problem that cannot be dealt with through regulations, because states and local districts do not have the capability for solving the problem. In the case of annual evaluation reports, a substantial part of the problem is that data are presented in such varied forms that aggregation across states to form a national program assessment has been impossible; technical assistance has aimed to promote uniformity of reporting, among other things.

   The information needed in order to implement a technical assistance program includes not only evidence of a problem but also information concerning proper methods for carrying out processes, and this information need requires research and development efforts that go beyond the usual type of evaluative information gathering. The area in which there is greatest need for technical assistance within Title I is the specification of effective methods for compensatory instruction, and in order to provide this assistance USOE has undertaken, among other things, to discover effective methods.

The many decisions involving actual delivery of compensatory education are made at the local level. The participation of state education agencies in the decision-making process varies greatly among the states and contributes to the local decision-making effort.

8. Local school administrators decide upon particular expenditures of Title I funds. The rationale for a choice among alternative projects would include information concerning which methods will generate the greatest reduction in educational disadvantage in the context of the local schools. Two forms of this information are (a) the results of careful research on compensatory education coupled with knowledge about the effects of the special context of the local district on compensatory education effectiveness or (b) finding that the methods used previously in the district's schools were satisfactory according to local standards. It is the purpose of local evaluations to provide the latter type of information; the general lack of valid evidence of effectiveness of locally developed methods provides the justification ior technical assistance from the federal government in the form of disseminating information about effective methods.

9. The teacher of a compensatory education participant, besides designating him/her for participation, makes day-to-day decisions on the form and content of compensatory instruction that for the child are at le st as important as any other decisions made in the system. Although these decisions have their rationales, the rationales are most frequently not clearly understood. It is an objective of curriculum packages to provide the decision rules (for example, in individualized instruction) that will enhance the child's achievement. Those decision rules are (ideally) the result of validation of rationales based on student performance during the development of the curriculum packages.

Although many other decisions might be included, these nine provide a basis for the specification of the prim ry information needs for Title I decision-making. Although it is the purpose of evaluation, generally, to meet these information needs, any particular evaluation project will meet only one or a few of the needs. An overall strategy is needed that would meet all the needs efficiently. The collection of information need not be related to decisions in a one-to-one fashion; not only are many decisions made simultaneously or in overlapping time periods, but certain types of information call for similar evaluation paradigms, some for different paradigms.

Information Required in Title I Evaluations

At the inception of Title I, information needs had not been clearly differentiated, and information gathering efforts designed to satisfy imprecise forms of all information needs at once were undertaken. As reported by Zimiles (1970), Wargo, Tallmadge, Michaels, Lipe, and Morris (1972), and McLaughlin (1975), the first five years of evaluation of Title I were essentially a total loss in terms of achieving any of the valid objectives for evaluation. In recent years, there has been greater differentiation of roles and objectives within the federal educational evaluation bureaucracy, and efforts such as the Descriptive Study of Compensatory Reading Programs (Trismen, Waller, and Wilder, 1975), the PIPs dissemination strategy (Stearns, 1977), the technical assistance centers and evaluation packages to help states and local districts carry out evaluations (Wood, Cannara, Fagan, and Tallmadge, 1976), and currently ongoing efforts funded through the Office of Education (the Sustaining Effects Study, System Development Corporation, 1976) and the National Institute of Education (the overall Title I assessment, National Institute of Education, 1976) are evidence of movement towards more realistic relations between objectives and operations in evaluations.

There are seven basic categories of information needed to test the rationales listed above. Various combinations of two and three categories of information, when properly analyzed, yield the required tests. The relations of the seven categories and their combinations to the rationales are shown in Table 1. Information on target and participant populations and on costs is "context" information; information on resource allocation is "input" information; information on management and services is "process" information; and information on effectiveness is "product" information. While the reader may disagree with some of the specific entries in this table, the important point is that such a relational table is a proper foundation for the development of a comprehensive evaluation strategy. Understanding how the information is to be used provides an important input to choices of ways of gathering the information (e.g., what populations the sample must represent and what particular details of information should be included in measurement instruments).

In order to provide this foundation, it is necessary to address four "systemic" questions, questions that concern the principles of the system's operat_or These may be addressed either as part of an evaluation project

Table 1

Categories of Information Required of Title I Evaluations

Category of Information	Needed to Test Rationale*
1. Target Population (level and frequency of needs; other characteristics)	2b, 3b
and Participant Population	1a, 2a, 3a
and Participant Population and Allocation Process	3a
and Costs	1a, 3a
and Effectiveness	7
2. Participant Population (numbers, per-pupil allocations, other characteristics)	
and Services ani Effectiveness	7
and Costs and Effectiveness	1b
3. Resource Allocation Process (selection of participants)	3a, 6
4. Local School Management Process (parental involvement, evaluation, project design)	6, 7
and Effectiveness	4, 5, 6
5. Services (processes, agents, contents, settings)	
and Costs	1a, 8
and Costs and Effectiveness	7, 8
and Effectiveness	4, 5, 7, 8, 9
6. Costs (resources needed for delivering compensatory education)	1c
and Effectiveness	1b, 4, 5
7. Effectiveness (changes in pupils' school performance)	8

*The rationales are numbered to match the presentation in the text. For example, "2b" refers to Rationale b for Decision 2.

or as a precondition to the design of evaluation projects. These four questions are:

1. What operations are intended to occur in the Title I system and how do they interrelate? In order that information gathered be relevant to decision-making, there must be a clear understanding of how the system is supposed to function, in greater detail than expressed in the law. For example, the meaning of "economic disadvantage," "evaluation," and "supplementary services" must be translated into specific observable events, if empirical observations are to be related to the program's principles.

2. What assumptions about society and human behavior are incorporated into the Title I system? For example, there would appear to be an assumption that economic disadvantage is a source of problems in schools that money can remedy. There also appears to be an assumption that children, once brought up to the ability levels of their classmates, will benefit from regular school instruction as much as their peers. Such assumptions must be separated from hypotheses about process-effectiveness in order that evaluation outcomes can be interpreted appropriately. In other words, the testing of rationales for decision-making should be undertaken with a clear awareness of the presuppositions inherent in those rationales.

3. What are the objectives of the program? For example, there needs to be a clarification of the types of impact on students that are to be considered as justifying Title I expenditures. Do these include cognitive skills beyond reading? Do they include attitudes and self-concepts? Do they include the physical well-being of the student? As another example, there needs to be clarification of the intended impact of Title I on the administration of local school districts. Should it include generally greater emphasis on promoting equality among all students or greater emphasis on evaluation and planning in school programs or more careful diagnosis of individual students' special needs? Also, to what extent is the objective of the program the mere transfer of funds to impoverished school districts? Mistaken assumptions about a program's real objectives will lead to useless recommendations; an

evaluator who understands these objectives can provide a more profound interpretation of his/her data.

4. What are the relative values of different program outcomes? This is a refinement and quantification of the preceding question. Not only does the range of objectives need to be identified, but also there must be some estimate of the relative importance of different outcomes. For example, to decide what percentage of a district's Title I funds should be spent on students in grades 1 through 3, it is useful to have some estimate of the value of compensatory education for children of different ages, based on a comprehensive theory of education. Likewise, for an evaluator to compare the benefits of different projects that achieve different goals, a quantitative measure of those achievements is necessary.

An argument against including systemic questions in an evaluation framework is that they are beyond the province of evaluators and are to be decided through political negotiation, logic, and common sense. As the work of many psychologists has shown, however, these processes are themselves subject to principles of human behavior that can be studied and improved upon. The use of scaling techniques to arrive at consensus values and the use of the research literature on social processes and human learning to identify the assumptions in the system are two instances in which systemic studies might well supplement the often bias-laden human processes such as political negotiation. Edwards, Guttentag, and Snapper (1975) have elaborated specific methods for dealing with some systemic questions in evaluation.

The general arguments for including systemic questions in the evaluation framework are (1) that otherwise they quite likely are not answered and the meaningfulness of the tests of decision rationales is therefore severely reduced, and (2) answers to systemic questions are more likely to represent the views of society at large if arrived at through systematic, replicable (i.e., scientific) methods.

Specific arguments can be made against forcing an evaluation to characterize the system in terms of a single set of objectives and outcome values. For one thing, the resulting set would oversimplify the situation. An important aspect of Title I is the multiplicity of goals of the program as viewed by citizens in different situations. By not delineating the operational objectives of the program precisely, Congress can forge a coalition of constituencies in favor of the program that might collapse if all the objectives were well specified. Any method of establishing objectives and values must not have the effect of forcing a collapse of the coalition. This is not an insurmountable barrier to rational decision-making, however. Systematically establishing the value dimensions for various outcomes of Title I would provide a much needed foundation for addressing fundamental evaluation issues, such as how to scale and aggregate achievement gains.

Information Gathering Processes

The aspect of rational decision-making most frequently referred to as "evaluation" is the gathering of information, or as it has recently been called, "the production of knowledge." For most evaluation specialists, the area of their training and technological expertise is in gathering reliable and valid information, and the choice among evaluators for a particular project usually depends on the demonstration of that expertise. Although evaluators are wise to be aware of the points discussed in the preceding section (i.e., how the information they gather is to be used), their primary responsibility is for gathering the information. In keeping with this concept of evaluation, the methodological issues discussed in this document relate primarily to information gathering, although the context of the information gathering will be seen to modulate the issues. (One general heuristic for this is that the evaluation of information gathering, like any other activity, should take into consideration the objectives of that activity. Another is that whenever you find you cannot gather a particular type of information, you should ask whether you really need it.

As set forth in Figure 1, we can view information gathering as consisting of four components: design, sampling, measurement, and analysis. The issues addressed in the subsequent sections are, in fact, grouped according to these categories.

The first of the components, design, is the most difficult to delimit. It is the planning process, the development or selection of a framework for information gathering. Thus, it overlaps the other three components: the detailed specification of the sampling, measurement, and analyses components would in fact include the total content of the design of information gathering. There are, however, three design factors that transcend the other components: (1) the general frame of reference, (2) the specific design model, and (3) the longitudinality of measurement. Much of the controversy concerning Title I evaluations has centered on the specific design models used. In particular, comparisons of performance of nonequivalent groups have proven faulty. The first design issue to be discussed will consider a possible resolution of that controversy by way of changing the evaluation frame of reference. The other design issue to be discussed pertains to the validity of evaluations based on gains within a single school year, a question of longitudinality.

The second component, sampling, refers to the specification of rules for selecting which states, districts, schools, projects, classrooms, or children to collect data from in order to reach general conclusions. Of the two sampling issues discussed, the first will focus on the impact of having nonrepresentative samples on the validity of the information provided, and the second will focus on the necessary size of samples to be used in evaluation.

The third component, measurement, has also been a center of controversy. There are three factors in the specification of measurements in evaluations: (1) the selection of which constructs to measure, (2) the selection, or development, of instruments (e.g., achievement tests) to make the measurements, and (3) the scoring, or scaling, of responses on the measuring instrument. Pertaining to the first factor are issues of how general the achievement gains are to be. These issues border on substantive issues of what the objectives of Title I should be; however, they also involve methodological issues of how to measure cognitive growth. Pertaining to the second factor is the issue of the role of criterion- and norm-referenced tests in evaluation of compensatory education, and pertaining to the third factor is the pervasive issue of the units of measurement, in particular, the role of grade-equivalent scores.

The fourth component of information gathering is analysis. Analysis has as its purpose the transformation of measures of sampled individuals into information relevant to rationales for decision-making. More particularly, the results of analysis are assignments of the likely truth of particular statements that contribute to rationales (e.g., "the likelihood that children would have learned this much in the absence of the program is less than 1 in 100"). The most salient issues concerning analysis are (1) whether there are adequate analytical methods for comparison between nonequivalent groups, and in particular, whether variants of analysis of covariance are appropriate; (2) whether analyses used to determine relations between effectiveness and costs and services are appropriate; and (3) whether methods used to aggregate data from different sources (e.g., annual state reports) are appropriate.

Realities of Evaluation

An important characteristic of evaluation is that it is a process involving people, and therefore the assumption that in reality evaluation conforms to some simple model, such as is presented here, will miss a large part of the true nature of evaluation. First, the purpose for which information is gathered does not directly affect whether it is reliable or valid; and large numbers of studies in the research literature are subject to the same methodological criticisms that are directed at federal evaluations of compensatory education. There are, however, two fundamental reasons why methodological problems appear to be more prevalent in the federal evaluation studies than elsewhere: (1) the results of the studies are of substantial importance to the lives of many people, so they are subjected to more intense scrutiny than are less sensitive research projects; and (2) requirements and constraints on information gathering are to a great extent specified by individuals with expertise in the use of information in decision rationales but not in the process of gathering information, and as a result the information gathering designs allowed are often limited to those of questionable validity (e.g., quasiexperimental designs; see Campbell and Stanely, 1963, and Campbell and Boruch, 1975). Only when policy-makers are aware of the alternatives for reliable and valid information gathering and of the consequences of basing rationales on less than adequate information can there be adequate evaluation.

One particular way in which evaluations of federal educational programs are limited is in their effects on the schools in which they collect data. Teachers and local school administrators naturally evaluate the goals of national evaluations as of secondary importance to the main task of teaching their students; and because the operations of information gathering do conflict with normal classroom activities, compromises must be made in order for any information to be gathered. One direction for creative solutions to methodological problems in evaluation may be in the negotiation of new forms of compromise between educators and evaluators.

Finally, before turning to the methodological issues, we should point out that the decision-oriented framework for evaluation that has been presented is not the only framework for evaluation. As pointed out by Floden and Weiner (1976), evaluations frequently have non-decision-making goals that complicate the identification of the information needed. "Evaluations" may be undertaken as a means of stimulating a project to take action, or as a form of public relations, or as a way of justifying decisions already made, or as a strategy in the development of an organizational power structure. While these activities are in a sense demeaning for evaluators who take pride in their information gathering craft, they nevertheless provide opportunities for the practice and enhancement of their craft. The methodological issues to be discussed are relevant to these activities also, to the extent that the information gathered might also be useful in future decisions; and, moreover, they are likely to be especially difficult to deal with because the individuals allocating resources to the "evaluation" are not motivated primarily to obtain reliable and valid information.

Summary

Evaluation of federal programs such as Title I has become a large-scale activity. Limitations on allowable information gathering continue to plague evaluators, however. This document will attempt to clarify the effects those limitations have on the validity of information gathered and to suggest potential directions for searching for solutions. Evaluation is viewed here as the gathering of information to test rationales for decisions, although the issues to be addressed are also relevant to other information gathering, or knowledge production, efforts. The information needs for Title I evaluation can be generally derived from consideration of the types of decisions Congress, USOE, local school administrators, and teachers must make, and they fall into seven categories: target population, participant population, resource allocation process, local school management process, services, costs, and effectiveness. In order to relate such information to decision-making, on the other hand, it is necessary to answer several systemic questions about the Title I system, either as a precondition or as a part of evaluation. Information gathering can be divided into four categories: design, sampling, measurement, and analysis; these categories provide the organization of the rest of this document. However, the issues to be discussed will involve political realities of evaluation that transcerd those four categories.

Design

Introduction

The design of information gathering is a dangerous topic for discussion. Many policy-makers view the technical aspects of it to be details that technicians can carry out, and they are concerned only with more global design issues; and many technicians view these technical details as the entirety of the design problem and fail to consider the more global design issues. Neither approach is satisfact‿y, however; the global issues and the technical details are actually closely interrelated. The best solution to a technical problem may be a change in the global design rather than an increase in the sophistication of techniques. Such a solution is proposed in the first issue to be considered in this section (the issue of the role of "control groups" in Title I evaluations). Although that issue has been considered by many researchers as a specific design procedure requiring further methodological development, a promising avenue for resolving the issue may lie in changing the frame of reference for evaluation. Policy-makers must listen to the expert advice of researchers and call upon other researchers to question and refine evaluation designs, if evaluations are to make use of the recent developments in methodology. Especially in the case of the first issue, the methodological sophistication in the research community in 1977 is significantly greater than a decade, or even five years, ago. Methods long accepted throughout the research community have been found questionable, at least as thev apply to evaluations of compensatoiy education.

The basic design problem, which has long been noted by philosophers of science (e.g., Eddington, 1958), is that all scientific observation and interpretation is perfor-ed in the context of a theoretical framework. Acceptance of the information thus gathered often depends on the acceptance of the framework. In particular, the use of statistical methods, such as estimation of effects in the population from effects observed on a sample, is predicated on sets of assump..ons that are rarely tested in evaluation studies. One reason for this is that statisticians have demonstrated that many of the most common methods are quite "robust" with respect to some of their assumptions; that is, the metho s would produce valid results even when the assumptions were not quite true. For example, the common t-test* assumes that random errors are normally distributed, but the test is quite valid even for significant departures from normality.

In the case f compensatory education, a recurring issue has concerned the validity of comparing achievement gains of Title I participants with the gains of a control group or of a standardized population. Methods for comparison of two groups in a psychological experiment are based on assumptions likely to hold true in the laboratory, but violated in the conduct of uncontrolled studies of ongoing programs in the field. One direction of resolution of these problems has been the "improvement" of statistical methods so that they involve fewer assumptions to be tested. That process is incremental; however, it is more costly than is generally recognized, and quite often it has taken the form of replacing one set of assumptions to be tested with another. The last deserves comment: it certainly is an advance to have two analytical methods that work for two different sets of assumptions rather than a single method; however, in practice having two frameworks requires the collection of extra information to test which framework is appropriate, which increases the overt cost of an evaluation. Plans for evaluations should explicitly include the assumptions underlying the observation and interpretation process and insofar as possible include plans for testing the assumptions.

*Student's t-test is a method for testing whether one group's scores are generally higher than another's. To carry out the test, one divides the difference between the group means by an estimate of how variable the scores are within each group. The larger the quotient, the more statistically significant is the difference in the group's scores. The aim of this monograph is not to serve as a statistical text, so particular statistical methods will only be described in sufficient detail to permit non-statistically trained readers to follow the discussion. The basic concept involved is that the truth of a statement is a function of the relative likelihood of obtaining a particular set of scores if the statement were true or were false. In the case of the t-test, it is the likelihood of obtaining a particular difference between groups if the difference were real or merely a chance occurrence.

In addition to the general question of what type of comparison should be used in Title I evaluations, there is another question of frame of reference that must be considered in planning an evaluation: What process is to be evaluated? Although, from a strict program evaluation perspective, it is the "Title I process" of allocating national resources to meet the special educational needs of disadvantaged children, the testing of rationales for decisions may depend more on information about other processes, such as "compensatory education," however funded, or "individualized instruction," or the relationship between economic and education-disadvantage. A problem in trying to evaluate the Title I process per se is the ability to separate those processes that have Title I as a cause from other processes occurring in the same classroom. Although superficially it appears that local school administrators are able to allocate Title I funds to identifiable categories, classroom dynamics preclude measurement of overall effects (e.g., if the effect of a particular Title I project is to pull students out for special reading instruction, the side-effects of this on the students remaining in the regular classroom cannot be ignored in a comprehensive evaluation of that project). Because the issue of what process is to be evaluated is determined more by considerations of the use of information than by problems in the gathering of information, it will not be considered as a separate methodological issue in this presentation.

Questions concerning the specific experimental design for information gathering have centered on the use of quasi-experimental designs to substitute for randomized, or true experimental, designs. Design in this sense refers to the operationalization of tests of decision rationales in terms of numerical relations to be observed among measurements of subjects (e.g., children) and the specification of subject selection in a way that will make inferences from numerical relations to tests of rationales meaningful and valid. There are dozens of common "experimental designs" that evaluators can apply to the evaluation task, but each is based on implicit assumptions that should be tested. It appears that at present we may be in a position in which none of the known "experimental designs" (including quasi-experimental designs) are both politically acceptable and able to provide valid tests of important decision rationales in Title I. This is discussed in Issue 1.

Questions concerning the longitudinality needed in Title I evaluation have centered on the paradox that children in compensatory education programs tend to learn about as fast as children who have no educational disadvantage, when measured from fall to spring; yet from year to year the children in compensatory education programs fall further and further behind their peers. Various aspects of this question are discussed in Issue 2 in this section.

The two issues discussed in this section by no means exhaust the methodological questions to be addressed in designing an evaluation study. They both focus on evaluations that aim to assess impact on students' performance. In addition to sampling, measurement, and analysis issues discussed in later sections, there are numerous "details" of procedure to which the director of a large-scale program evaluation must attend, such as staff management, scheduling, liaison with program and project managers, data management, and report design. An evaluation is equally susceptible to loss of credibility from carelessness in these aspects as from statistical design problems.

Issue 1. To what should Title I treatments be compared?*

Evaluation is not mere description. Where description is substituted for evaluation, important systemic questions about the program have not been addressed. The testing of decision rationales through the gathering of information always involves the interpretation of that information as it relates to a hypothesized description of the program. Thus, there must always be some comparison of the information on program performance, in our present case the achievement of children who have received Title I services, with a standard. Based on this comparison, the validity of decision rationales can be tested, and recommendations for policy can be developed. As pointed out by Stake (1967), there are two fundamental types of standards for comparison: (1) comparison with what would have occurred were the treatment not present-- a relative comparison; and (2) comparison with a goal-outcome that the treatment

*A summary of the types of treatments funded under Title I is contained in a companion volume (McLaughlin, 1977).

was intended to produce--ar absolute comparison.*

In the case of relative comparison, the estimation of what would have occurred without the special treatment is the most significant design problem; in the case of absolute comparison, the specification of the goal-outcome desired is the most significant design problem. Methodologies for estimating "what would have occurred" are noticeably further developed than the (more complex) science of educational goal-setting (e.g., the t-test is universally accepted, but goals for the schools vary from one community to another). For that reason, it would seem, compensatory education evaluations have been designed for relative comparisons. The recent development of "criterion-referenced tests," "objective-referenced curricula," and "competency-based education" (Spady, 1977) is, perhaps, a harbinger of a movement toward specification of goal-outcomes for compensatory education, which would allow absolute comparisons. Both types of comparison play a role in ideal program development, as shown in Table 2. They are based on distinctly different points of view, however. The type of question answered by a relative comparison is "Did the program have an effect?", and the type of question answered by an absolute comparison is "Did the program meet the need?" A relative comparison will not tell us whether the problem is being solved by the treatment, and an absolute comparison will not tell us whether the level of expenditure is justified. A comprehensive evaluation strategy would require both types of comparison.

In order to resolve this issue, it is necessary to weigh the costs and benefits of the various alternatives for comparing Title I treatments. We shall first consider the intricacies that have been discovered in using relative comparisons and then examine the potential for the use of absolute comparisons, which have received little attention in the ten years of Title I evaluation.

*Lest the terms "relative" and "absolute" confuse the reader, it should be noted that in a sense all comparisons are relative. The way these terms are being used here refers to the dependency of the validity of the decision rationale being tested on the operationalization of some hypothetical model (e.g., what would have occurred in the absence of Title I). A "relative" comparison is so dependent, and an "absolute" comparison is not.

Table 2

Relationships between Outcomes
of Absolute a d Relative Comparisons

		Absolute Comparison	
		Positive Results	Negative Results
Relative Comparison	Positive Results	Conclusion: Program operation satisfactory.	Conclusion: Need for more effort or reconsideration of goal-outcomes.
	Negative Results	Conclusion: Program effort can be decreased substantially or goal-outcomes need reconsideration.	Conclusion: Need for redirection of program efforts and need for new methods.

Relative comparisons. For relative comparisons, the only method known for estimating what the perfoimance level would have been without the treatment is to observe the performance level of some other group not receiving the treatment*. The selection of that other group is crucial to interpretation of the comparison. There are three categories of alternatives: (1) random assignment of preselected subjects to treatment and no-treatment (i.e., standard school treatment) conditions; (2) selection of a comparison sample in any other way; and (3) use of norms tables of estimated performance in the general population, published with standardized tests. Random assignment is necessary for the true experimental method; it involves the least threat to the internal validity of evaluation but the greatest complexity in interaction with program operation. Random assignment, it should be noted, can refer to assignment of students within a classroom to treatment and control groups, to assignment of schools to treatment and control conditions, or any other unit. The implications of randomization of different levels of units are discussed under Issue 3. The only major federal education program evaluations that have employed randomization are the ESAP and ESAA evaluations (NORC, 1973; Coulson et al., 1976).

Nonrandom comparison groups that have been used in major Title I evaluations include (1) students in the same school in a prior year (Mosbaek, 1968), (2) students whose classmates were participants in Title I (Trismen et al., 1976), and (3) students without compensatory education programs (Trismen et al., 1976). Nonrandom comparison groups have been used in numerous local evaluations, and current efforts by USOE to help local districts carry out Title I evaluations include this method (Wood et al., 1976). The problem with nonrandom comparison groups is that there is no assurance that they are comparable to the treatment group prior to treatment. As we shall see, the ways in which they can differ are numerous, and testing for all the possible differences

*A repeated measures design in which each child acts as his/her own control is an interesting alternative, but would involve extremely complex corrections because the goal of compensatory, and regular, education is to change the child, and the rate of individual growth varies in complex patterns from year to year.

so one can make the correct adjustment of the comparison borders on the infeasible. On the reverse side of the coin, there is, by the fact that Title I is designed for a subset of the children in a school, nearly always some comparison group nearby that can be inexpensively tested to provide comparison data. One type of comparison data available in many school districts is cumulative growth curves for children in the district. Although subject to problems, these local norm data are usually preferable to the use of national norms.

The use of national standardization data has much the same set of problems as use of a nonrandom comparison group. The problems are compounded by the fact that, unlike a contemporary local comparison group, one cannot observe what variety of experiences and traits characterize the national comparison group. The problems associated with use of the norms tables of standardized tests are discussed under Issue 6. Nevertheless, such data, in the form of gains relative to typical performance at a grade level (grade-equivalents), have been used by many states for their annual Title I evaluation reports and thus by the federal evaluators who aggregated the state reports. Of course, for the local evaluator, use of norms tables is the least expensive method for generating a comparison of a treatment group's performance.

In order to evaluate the usefulness of these three methods for performing relative comparisons, we must consider the various costs generated by each alternative. Four types of marginal costs must be included:

1. costs of collecting the needed data for comparison;
2. costs of producing the data (incurred by teachers and students);
3. costs in lost validity and in resulting lost credibility of the evaluation's findings, compared with other methods; and
4. costs for development of the method.

These costs offset each other, and they apply to absolute comparisons as well as to relative comparisons. Therefore, it is essential for an evaluation designer to understand their differences and to be able to compare their values. Because the credibility of the findings is partially dependent on what the findings are (that is, whether or not they conform to results desired by groups in a position to attack their validity), policy-makers and evaluation designers must choose whether to gamble with an "inexpensive" design and hope that results prove noncontroversial or to be conservative and use an "expensive" design. A pilot evaluation is a useful tool in this situation.

For random selection, marginal costs are nearly all in the category of producing the data. The major cost, invariably given as the reason for ruling out randomization, is the withholding of Title I benefits from the unlucky needy students selected to be in the control group. The law specifies that Title I funds are to be used to meet special educational needs of the young people with the greatest needs, and program administrators and teachers are reluctant to compromise that principle merely for the purposes of valid evaluation. This is a constraint within which evaluation must be carried out. Proponents of randomized designs must find rationales for randomization that will meet the objections of administrators and teachers.

Several such rationales have been suggested (e.g., Campbell and Boruch, 1975). First, one might argue that there is little evidence that missing out on the program for a year has lasting effects on one's education; after all, "no Title I treatment" does not mean "no instruction." Finding that local districts, teachers, and parents do not readily accept this argument would indicate by itself that these people, at least, believed the treatment to be effective.

A second design for randomization is conceivable when more than one compensatory service is available, only one of which a student can receive at a time. For example, if there are compensatory reading and mathematics classes, then it might be reasonable to assign needy students randomly first to one for a year and then to the other for a year. This would be questionable if children normally were behind in only one of the subjects: assigning a child with math difficulties to a compensatory reading class might be counterproductive. The results, at least for the first year, would be a randomized design in which each compensatory group was the control for the other. This presumes that the content of two instructions has little overlap (otherwise the evaluation would be too stringent, pitting two compensatory classes against each other); a presumption probably false in the primary grades. Very sensitive tests would be necessary to differentiate gains of two classes whose objectives overlap.

A third randomization design would be to withhold compensatory education service from a randomly selected group of stuaents for a year and invest the money saved in a trust fund for those students. Although this possibility is bizarre, it should not be dismissed without consideration. Perhaps the most difficult problem for this design is the fact that there are substantial side effects of the introduction of Title I funds into a school that would not be felt if the money were in the bank.

A fourth design for random-assignment can be used when there are not sufficient Title I funds to serve all the needy students. Rather than dilute the program's effectiveness by giving each student less service, and rather than assigning funds on some basis such as ability of a teacher or school administrator to write a good program proposal (which may not be indicative of the actual service delivered), some of the funds could be assigned randomly. This procedure is fair and can be agreed to in advance. Although it would be infeasible to implement at the level of selecting individual students, it proved feasible in the selection of schools for ESAA money in the evaluation designed by USOE and carried out by the System Development Corporation (1976). As that evaluation showed, however, it is necessary to have advance agreement that no compensating local resources that might affect the level of performance of students in the control schools will be allocated to those schools during the period of the evaluation. In that study, because the Office of General Council held that USOE administrators could not affect the allocation of other resources to make up for ESAA allocations, the evaluation was compromised. The general heuristic of substituting a service or value to be provided after the evaluation is completed appears to be a reasonable compromise between program operation and program evaluation.

A fifth possibility occurs in districts with a wide range of economic status, where it is required that local administrators select the schools serving the most economically disadvantaged children to receive Title I assistance and demonstrate that non-Title I schools are not receiving compensating resources from other sources. In these cases, random assignment of a few groups of children to Title I and non-Title I schools would provide the basis for an overall comparison between the Title I and non-Title I schools, although it would be difficult to make inferences about the effectiveness of particular methods in such a design.

Finally, if the base of comparison were taken not as between the Title I treatment of interest and the standard instructional treatment but rather between a Title I treatment of interest and a standard that is agreed to be highly effective (although possibly too costly for widespread use), then random assignment could easily be justified. The aim of this comparison would be to show whether the treatment of interest was as good as the "standard of excellence," presumably at less cost. This provides an argument for the identification of at least one method of compensatory education, however costly, that can be assumed successful wherever implemented.

To summarize, the costs of randomization are nearly all in terms of services withheld, and several rationales exist for compensating for or justifying the withholding of services. Of course, a thorough consideration of randomization would have to investigate secondary costs for the teacher and for other students: for example, the greater classroom homogeneity of achievement level when low achievers are taught separately might possibly benefit noncompensatory classes as well as the compensatory classes (although the results of Trismen et al., 1975, suggest not)--randomization removes that possibility. However, in view of the marginal costs of the other methods to be described, randomization deserves careful consideration (as in Conner, 1977) for future evaluations that require relative comparisons.

There is another cost associated with use of randomized control groups that applies equally to nonrandom comparison groups but not to other comparison methods. This is the cost of assuring that the control group is not affected by the Title I service; if it is affected, this would bias the comparison. There are numerous sources of subtle effects of which the evaluator must beware and which he/she must either avoid or measure and correct for. If students in both groups are in the same classroom or even the same school, some peer teaching of skills learned in the compensatory treatment will be very likely to affect other students; if a group of students is aware that they are being used as the control group, competitive spirit will lead to greater achievement than were there no evaluation (the "John Henry effect"); teachers are likely to discuss with each other methods that have been successful, thus spreading their use; if both groups are in the same classroom, the teacher may notice "mistaken" assignments to treatment and control groups and reassign students to achieve maximum benefit from the compensatory education program, ignoring the effect of this on evaluation; and districts may unconsciously favor schools not receiving Title I money with other opportunities "in order to be fair," although that is precluded by Title I regulations. Thus, randomization or other methods of selection of a control group will have costs associated with the proximity between treatment and control groups that other comparison methods do not.

We turn now to nonrandom comparison groups. The problems of evaluations involving nonrandom comparison groups have been discussed at greater length than any other methodological topic in the relevant literature (e.g., Thorndike, 1942; Campbell & Stanley, 1963; Campbell & Erlebacher, 1970; Glass, Peckham, & Sanders, 1972; Kenny, 1975; Porter & Chibucos, 1974; Sherwood, Morris, & Sherwood, 1975; Campbell & Boruch, 1975; Boruch, 1976; Reichardt, 1976.) Although we leave the details of the methods of analysis when one has nonrandom control groups to the discussion under Issue 8, we shall consider the problem generally here in order to understand the costs involved in choosing to use a nonrandom control group for a relative comparison in evaluation.

The basic problem is that the treatment and comparison groups must be determined to be equivalent in all relevant aspects, so that they can be compared "as if" the selection had been random. That equalization which is a form of interpretation of observations, depends on assumptions. Those assumptions are numerous, and testing them is both necessary and costly. While there have been notable advances in expanding the available methods for correcting for the nonequivalence of control groups, there have been equally notable additions, especially by Donald Campbell and his associates, to the list of problems that must be dealt with in analyzing data from "quasi-experiments" (as Campbell & Stanley, 1963, referred to designs without randomized assignment).

In order to understand the scope of the problem, let us consider the simplest form of correcting for the nonequivalence of control groups, one that has been berated often, is still often used, is really no worse than some more sophisticated methods, and one form of which was recently strongly defended (Sherwood, Morris, & Sherwood, 1975). This method is matching: for each treatment subject, a control subject is selected to be as similar as possible to him/her before the treatment, and differences are measured between the pairs on completion of the treatment. The following list of problems with this method, taken from Campbell & Boruch (1975), is incomplete, but will show the extent of the problem. It should be noted that the method proposed by Sherwood, Morris, & Sherwood (1975) may not be subject to many of these problems, because theirs was an attempt to match pairs exactly--on dozens of dimensions simultaneously. These problems listed are primarily for the case in which matching is on a pretest.

1. Differential regression to the mean: Children selected by their teachers as needing compensatory instruction are likely to have obtained low pretest scores because their true scores are low; however, those noncompensatory students whose low observed scores match the compensatory students are likely to have obtained the low scores through random error. On retesting, their scores would be expected to be higher than the matched compensatory students because the random error would not be likely to be in the same direction. The problem is that matching is on observed scores, not on (unobservable) true scores, and the result is that compensatory education can look bad entirely due to the statistical artifact. A hypothetical example is shown on the next page.

	Five Compensatory Education Students					Five Regular Students				
	A	B	C	D	E	F	G	H	I	J
Pretest True Score	20	20	30	40	40	30	30	40	50	50
Pretest Observed Score	15	25	30	35	45	25	35	40	45	55
Matched Students		25		35	45	25	35		45	
Posttest True Score if Everybody Gains 5 Units		25		45	45	35	35		55	
Posttest Observed Score		30		45	40	40	35		50	
Average Pretest Observed Score for Matched Groups			35					35		
Average Posttest Observed Score for Matched Groups			38.3					41.7		

2. Differential growth rates: Children who learn more slowly are farther behind their peer group at any age, and conversely, children who are farther behind tend to have a slower learning rate, at least in most cases. Two students may have achieved the same level, however, and have different growth rates, for example because the slower student was given extra help. Now, an intelligent teacher is likely to be able to discern which of two children scoring low on reading has a real learning problem requiring compensatory instruction and which is merely not performing up to his/her capabilities and can be expected to cope with the tasks in the regular class. It would surely be unfair to the compensatory treatment to match these two children for the purposes of evaluation. See the example below.

	Compensatory Education Student	Matched Control Student
Pretest True Score	20	20
Rate of Growth	10 points per year	20 points per year
Expected Posttest Score	30	40

3. <u>Test floor effects</u>: Each achievement test is designed for a particular range of ability levels. However, if the test used for an evaluation is not very carefully chosen, some low achievers may in fact have pretest true ability levels significantly below the level needed to exceed chance (pure guessing) performance scores. For example, in the Compensatory Reading Study (Trismen e [illegible], 1975), there were numerous means for groups of compensatory reading students that were below the guessing level for the test. The pretest scores of students who purely guess will be positive, however, because some guesses will be correct, and they will be matched by controls who perform at chance levels that reflect their true scores. In the course of a school year, the treatment and control students might learn an equal amount, but that amount might not be enough for the treatment students to exceed chance levels. Thus their observed gain would be zero, compared to a positive gain in the control group. See the example below.

	Compensatory Education Student	Matched Control Student
True Pretest Score	10	20
True Score Needed for Chance Level Performance	Anything up to 20	Anything up to 20
Observed Pretest Score	20	20
True Gain for Year	10 points	10 points
Observed Posttest Score	20	30

In each of these cases, a test could be made for whether the particular biasing effect actually occurred and a correction made. For example, parallel forms of the test could be given to each student to estimate the amount of regression to the mean, and scores could be corrected before matching. Measures of growth rate could be obtained by administering several pretests over a period of years preceding the treatment. Test floor effects can be avoided by pretesting the tests before using them for the evaluation study or by the development and use of wide range tests, such as the sequential branching tests that can be administered under computer control. Each of these operations adds significantly to the cost of the evaluation, however, and it is not too cynical to expect that a sophisticated methodologist will be able to find some new source of bias after the study is completed. In some cases, it may be expected that the results will be so clear-cut that statistics are hardly necessary (for example, if all students in some compensatory program scored in the bottom half of their class on the pretest and in the top half of their class on the posttest, no statistical artifacts would be important). It also may be that the results will be noncontroversial (for example, if they are merely to corroborate results obtained from different methods of evaluating the particular program). In these cases, the pressure on internal validity is not as great, and one might conclude that the cost in units of credibility may not justify abandoning the alternative of matching. The history of politicization and controversy of Title I evaluations, however, suggests caution in sacrificing validity to save other costs.

Another approach to this problem, which has its own costs, is to develop airtight methods for interpreting results based on nonrandomized studies. Porter (1967) and Kenny (1975), for example, have improved the methodology (to be discussed under Issue 8), and the National Science Foundation and the National Institute of Education have recently been supporting some research into better methods, so the possibility of the development of improved analysis procedures for noncomparable control groups should not be dismissed. The proper method is not apparent in 1977, however, and there is no guarantee of solution in the near future. Nevertheless, more intensive effort in this direction seems warranted, unless either randomization becomes politically acceptable or evaluations change toward absolute comparisons rather than relative comparisons.

The third method for relative comparisons is to compare the Title I participants with the "norm group," that is with the scores of the representative national sample of students who took the test before it was published in order to establish the meaning of the raw scores in terms of comparison to the population. The model used for such comparisons in an evaluation is the "equal growth" assumption. This is the assumption that, under no special treatment, a student who scores at, say, the 20th percentile relative to others at his grade level (or 7 months below grade level or 10 items or 1 standard deviation below the mean) at the beginning of one grade is expected to score at the 20th percentile relative to his peers (or 7 months below grade level or 10 items or 1 standard deviation below the mean) at the beginning of the next grade. All of the validity problems of nonrandom control groups apply equally to this method of comparison, and it also is subject to the numerous problems that arise from reliance on norms (see Issue 6). Moreover, Kaskowitz and Norwood (1977) have presented data that indicate that the equal percentile growth assumption leads to underestimation of expected gains of students at the lowest percentiles, based on data from recent evaluations; and the distortions associated with use of grade-equivalent scores are well-known (see Issue 7).

In view of the numerous problems associated with use of test norm data as a comparison standard for compensatory education evaluations, it is distressing to find that most evaluations carried out to satisfy the requirements of Title I have been based upon that type of data (see the discussions of local and state evaluation reports by Wargo et al., 1972; Gamel, Tallmadge, Wood, and Binkley, 1975; Thomas and Pelavin, 1976). The use of such data is even recommended as one alternative for future local Title I evaluations (Wood et al., 1976). Only when special research studies have been commissioned by the federal government and carried out by leading research institutes have there been comparisons with control groups (most notably the Compensatory Reading Study, Trismen et al., 1975, and the Sustaining Effects Study, System Development Corporation, 1976).

The cost of this method (norm comparisons) in terms of data collection is minimal, but from the point of view of validity it is substantial. For the purposes of relative comparison in evaluation, its use should be corroborative rather than as a sole means of comparison. The costs of development to establish adequate validity for this method include not only the costs associated with developing methods for interpreting comparisons with nonrandomized contrcl groups, but also they include the costs of refined standardization, in which distributions of scores in the norm sample are crosstabulated with numerous demographic and other factors that might be used to match the treatment group to a subset of the norm sample.

As we have seen, there are substantial problems to be dealt with in the use of any of these alternative methods for relative comparison. Any one of them might provide the answer: if a politically feasible method of randomization were developed, or if sufficient statistical methods for equating nonequivalent comparison groups were developed, or if sufficiently reliable and valid test norms were produced. The stakes are sufficiently important (Title I is spending about $2 billion annually and is substantially affecting the education of 5 million children annually) to warrant strong efforts in all three directions. It is our belief, however, that a fourth alternative, turning to absolute comparisons in the evaluation of Title I impact, is also viable, and we have taken the next few pages to discuss that alternative.

Absolute comparisons. Absolute comparisons involve comparison of a treatment group's performance with an agreed-upon standard, irrespective of any control group's performance or, really, of any form of expectation for the treatment group's performance. Four types of absolute comparison standards, shown schematically in Figure 2, appear to be reasonable for the evaluation of impact of Title I on children's educational attainment:

1. specified minimum skills to be achieved at each grade level;
2. specified maximum deficits from the population average to be allowed at each grade level;
3. specified minimum amounts of skill acquisition per year of school; and
4. specified minimum amounts of deficit reduction relative to the population per school year.

The first two standards are for achievement levels at the conclusion of

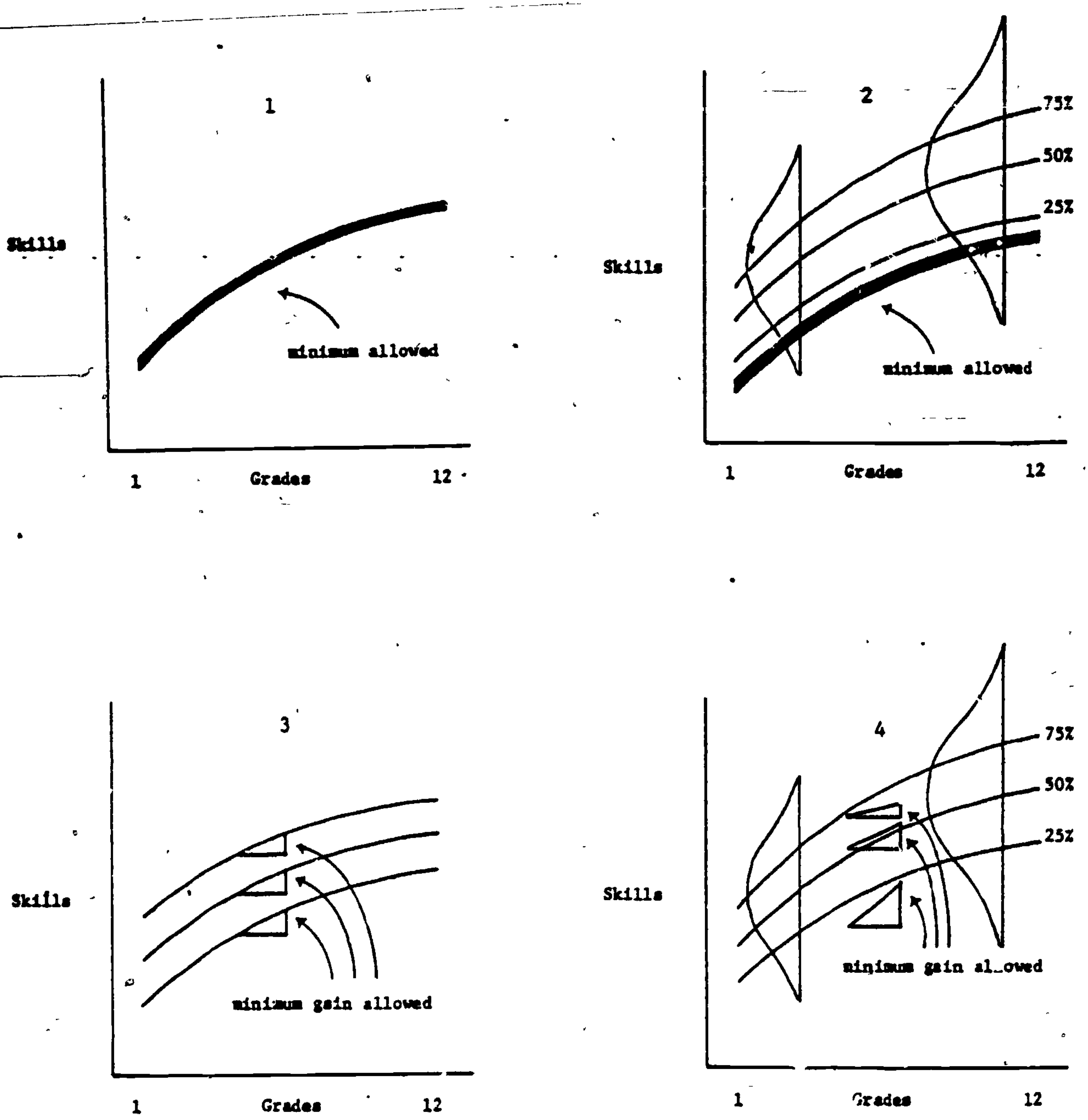


Figure 2. Four kinds of absolute comparison standards for Title I achievement gains

a particular Title I treatment, and the last two are for achievement gains. The first and third are in terms of absolute skill levels, and the second and fourth are in relation to what skills the population as a whole possesses. Although all but the first are expressed as the relationship of posttreatment performance relative to some other performance level (pretreatment or the general population), they are nevertheless absolute comparisons in that they can be agreed upon ahead of time, their validity in no way depends on the ability to find an equivalent control group with which to compare the treatment group. For example, of in the second type of comparison, the criterion for concluding that Title I is having the proper impact is that every participant's performance be at least at the 25th percentile of the population distribution upon completion of the treatment*, it is immaterial how the particular treatment group differed from typical students in the population prior to treatment.

As with the alternatives for relative comparisons, the types of costs for the four methods of absolute comparison vary, and careful analysis must precede selection of the appropriate method. The only applications of the methods to Title I evaluations have been in the searches for exemplary projects (Wargo, Campeau, and Tallmadge, 1971; Horst & Tallmadge, 1975), and a substantial amount of development will be necessary prior to their widespread use. Recognition of the need for such development is apparent from the attempt by Horst and Tallmadge (1975) to achieve a measure of what they termed "educational significance" in terms of a comparison of the fourth type. They proposed that in a search for successful projects one require not only that a gain be statistically significant, but also that the amount of the gains

*This is not paradoxical: it requires that the distribution of skills be truncated at the 25th percentile, so that the raw score for the 1st and 25th percentiles would be essentially equal.

be at least 1/3 population standard deviation*. Kaskowitz and Norwood (1977) have pointed out the need for improving on this arbitrary criterion before extending its use to other evaluations, implying that it will be extended whether it is refined or not. Horst (1977) has investigated the relationship of gains of 1/3 standard deviation in a school year to typical amounts learned in a year. He found that to gain 1/3 standard deviation a student who is one standard deviation below the mean in an early grade must learn twice as much as would otherwise be expected and a student in an upper grade must learn three or four times what is normally learned in a year.

Of the four types of absolute comparison, there is little difference in data collection cost: the only variation is that pretreatment performance levels must be obtained for the third and fourth methods in order to calculate gains at the time of posttreatment testing.

Costs for development and for credibility are interchangeable. With a minimal effort, experts could be brought together to draw up a tentative list of skills to be achieved at each grade level, for example, but selecting a single set of skills and gaining universal acceptance for it

*The population standard deviation is an estimate of how far one expects particular scores to be from the population mean on the average. For a normally distributed score, about 68% of the scores are within one standard deviation on either side of the mean, and about 28% more are between one and two standard deviations from the mean, as shown in Figure 3. A gain of 1/3 standard deviation for an individual at the 16th percentile, for example, would move that person to the 26th percentile.

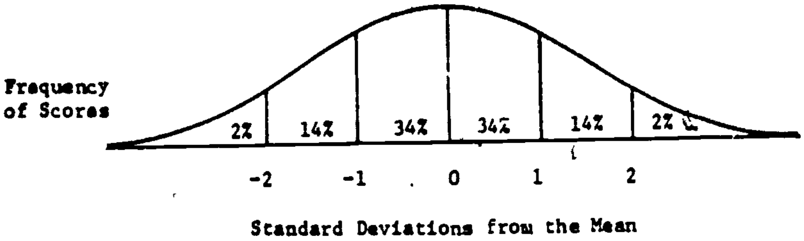


Figure 3. Normal distribution of scores

is a mind-boggling task. We will consider here the developmental costs in some detail for the first method because it is potentially the most far-reaching, and the costs for the other methods involve primarily subsets of the cost components for the first method.

First, a method for deriving minimum proficiency levels at each grade level must be agreed upon. Two alternatives present themselves. The first involves working backwards from minimum proficiency levels that are to be obtained by the end of 12 years of schooling. Oregon, California, Michigan, and a few other states are beginning to implement a policy of minimum proficiency testing for high school graduation, with each local school district developing local minimum standards. Spady (1977) has suggested that this will be a widespread practice in the near future. There is a significant problem in "working backwards" from exit-level requirements to requirements for each grade level, in that there are many alternative paths to the learning of basic skills. While there has been a great deal of research on the hierarchy of skills involved in reading (Williams, 1973), and the National Institute of Education has focused a large research effort on the process of learning to read, there has been no attempt to translate the results into a set of alternative paths toward minimum proficiency.

The second approach to establishing minimum levels at each grade is theoretically less ambitious and more appropriate to the basic assumption of Title I that compensatory education can bring students back into the mainstream where they can benefit from regular school instruction. This approach is to establish the skills necessary for benefiting from each particular classroom's regular instruction and set the goals of the previous grade's compensatory instruction to bring as many students as possible up to that level. This could be facilitated by curriculum developers' specifications of skills needed for their published materials. Of course, the cost of generating such specifications (correctly) would be quite significant, and whether that cost should be reflected in higher costs for textbooks or treated as a governmental responsibility is not clear.

Of these two approaches, the former has the advantage that it treats the schooling process as a single system that produces skills in students leaving the system sufficient for coping with life's problems. The latter has the advantage of more easily fitting into the existing education framework in each school district, but it requires that teachers in successive grades get together to set up objectives for compensatory education, and it ignores the needs of children who frequently switch schools. On the dimension of evaluation credibility as a tool for program development, the audience must be specified to decide between these approaches: for the local district, the second approach is more beneficial in that it facilitates incremental improvements within the existing framework. At the national level, where the concern is for preparation of the adult citizenry of the next generation, the results of the first approach are more meaningful. To arrive at a choice between these methods clearly requires additional work.

Turning now to the second type of absolute comparison (using a national average), theoretical problems of defining what skills are really necessary are replaced by the empirical problem of determining the percentages of children at each grade level possessing various cognitive skills and by the systemic problem of determining hwo close to "equality" of achievement to aim for. Of course, absolute equality of achievement is an unattainable and indeed undesirable goal in a free pluralistic society.

One answer to the question of "how equal" the program should aim students is to use data (for example, from the National Assessment of Educational Progress) to estimate the percentage of young adults nationwide who do not possess minimum proficiency levels agreed upon by experts and, after correcting for various statistical artifacts such as varying rates of early dropping out of school, set that percentage as the goal for compensatory education. For example, if it is determined that 15% of the young adult population is mathematically incompetent upon high school graduation, this means that 85% are judged at least minimally competent (i.e., not requiring federal intervention). Roughly, this implies that if the performance of all students at each grade level is maintained at levels within what is the top 85% (above the 15th percentile) of the existing population, all students graduating from the system in the future will be mathematically competent (by standards of the 1960s and 1970s)*.

Contrasting the first two types of absolute comparison, one arrives at the conclusion that the decision between these two methods for evaluation depends on the answers to crucial systemic questions about the role of Title I in society: is it to ensure a certain minimum skill level or to ensure a certain approximation to equality of achievement?

The third and fourth types of absolute comparison both differ from the first two types only by taking into account the students' levels of performance at the beginning of participation in a Title I program. We can, therefore, discuss them as one. The primary advantage of expressing goals in terms of gains rather than absolute levels is that failure to meet the criteria can more easily be attributed to deficiencies in the program: of two programs evaluated completely in terms of posttests, one might appear more successful merely because its students were further advanced at the beginning of the program. It would be wrong to select programs on that basis. The main drawback of using gains as the criterion is that they do not relate directly to practical criteria, such as possession of particular skills after 12 years of school, possession of skills necessary for regular instruction in the next grade, or performance at a specified level relative to the population. If a student is sufficiently far behind upon entry, then even extraordinary gains may leave him/her below desired posttreatment levels. One way in which criteria could encompass both concepts (using gains and relating to absolute posttreatment levels) is to specify that gains should be enough to close the gap between pretreatment levels and desired levels by a

*The reader should not fall into the trap of worrying that there will always be a bottom 15%. Of course there will; however, the goal would be for their skills to be above what is now the 15th percentile. As society changes in the 1980s and 1990s that criterion could be expected to change.

significant fraction, such as halfway, if the student has more than a particular specified deficit on entry. A more thorough solution would be to perform both types of absolute comparison (i.e., with and without correction for pretreatment performance levels). The implications for policy are shown in Table 3; they are analogous to the differential implications from absolute and relative comparisons shown in Table 2.

Summary. The problem of what to compare Title I treatments to is complex and involves careful analysis of the program's basic assumptions. We have considered seven classes of alternatives: relative comparison using randomized assignment, nonrandom comparison groups, and national norms as standards, and absolute comparisons involving either posttreatment levels or gains and either prima facie skill requirement specification or specification in terms of the skill level in the society. We have not considered any number of other dimensions that should be in a more thorough treatment of this subject: Are comparisons with, say, Great Britain relevant? Are comparisons with the society's costs of dealing with functionally illiterate adults relevant? Are comparisons with state compensatory education programs relevant? One can certainly imagine rationales for important decisions that would depend, at least ~~partly~~, on the answers to these questions.

Although it is impossible to rule out as inappropriate any of the seven categories of comparison discussed, it appears that much of the lack of direct impact of evaluations on program operation (Cohen & Garet, 1975) may be due to complete focus on relative comparisons, which merely test whether a program is better than what was being done previously, rather than absolute comparisons of whether the program is achieving specific educational goals. The change of focus toward absolute evaluations is needed and shows signs of occurring in the near future.

Table 3

Implications of Annual Comparisons of Gains or Posttreatment Levels

		Comparison of Posttreatment Levels	
		Positive Results	Negative Results
Comparison of Gains	Positive Results	Conclusion: Program operation satisfactory.	Conclusion: Greater program effort needed; or gain criterion needs revision; or posttreatment comparisons should await another year's gains
	Negative Results	Conclusion: Program effort not dealing with a clear need.	Conclusion: Program needs redirection, new methods.

Issue 2. Is longitudinal evaluation necessary?

A longitudinal design is one which requires the collection of data from the same source two or more times over a period of time. In order to determine whether a longitudinal design is necessary for Title I evaluation, it is necessary to examine the information needs guiding the evaluation to determine whether they warrant the expenditure of effort required for longitudinal data collection. The first question is whether temporal relational information is necessary. If it is, the next question is whether less problematic designs, retrospective data collection or cross-sectional designs, would provide sufficiently valid information. The final question is, if temporal relational information is required, over how long a period of time must the data be collected.

The answer to the first question is that for evaluation of impact on children's school achievement, although not so clearly for gathering information on compensatory education processes, temporal relational information is likely to be necessary. As long as the impact is measured in terms of gains, starting from disadvantage, there must be some version of a "before" and "after" measure. If the framework of comparison were to be oriented to the comparison of posttreatment levels with a standard, irrespective of pretreatment differences, temporal relational information would not be so important; however, that would require a substantial break from the current evaluation tradition.

The temporal relational information normally required has three components: (1) a measure of a child's achievement level prior to the Title I treatment, (2) a measure of the child's participation in the treatment, and (3) a measure of the child's achievement level following the treatment. The second question we posed for deciding on longitudinal designs was whether the information could be gathered by easier methods. The easiest would be retrospective data collection, use of a respondent's memory to construct temporal relational information; however, that is not feasible for the assessment of pretreatment achievement levels. As survey methodologists have frequently pointed out, the reconstruction of previous subjective variables has little validity, and retrospective questions should be limited to questions concerning actual, objective events, such as switching of schools, participation in special classes, and sc on. This lack of validity of retrospective subjective reports applies also to the reports of teachers that (although the test scores may not show it) the children in their classes improved "significantly" during participation in a program (e.g., Stearns, 1977). Although teachers may be quite sincere in these reports, there is a great likelihood that they may be based on the teachers' unconscious selective perception of behaviors that matched their expectations or desires.

The second alternative is a cross-sectional design. Rather than collecting pretreatment and posttreatment scores on the same students, it might suffice to collect them on different students. The reasons for doing this might be (a) to circumvent the methodological problem with pretest-treatment-posttest designs that the pretest may itself affect the way in which the treatment is perceived and assimilated by students, or (b) to shorten the time needed to study a long-term treatment (e.g., one could estimate 4-year gains by measuring 2nd and 6th graders in a school at the same time).

The primary requirement for inferring temporal relational information from cross-sectional designs is that the samples on which the different measurements are made be equivalent in all relevant respects. For cross-sectional designs aimed at the first of the two problems (effects from the pretest), this can be accomplished by randomly assigning students to either a pretest-treatment or a treatment-posttest condition or, better yet, by randomly pretesting only half of the students, posttesting all of them, and testing for the existence of a pretest-treatment interaction. There has been little, if any, use of such a design in Title I evaluations to avoid pretest-treatment interactions, probably because of other advantages, to be listed below, of true longitudinal designs. One particular problem that has been rarely recognized is that, in comparing a treatment's gains with a national test norm, the students are normally taking the test (a parallel form) for the second time at the posttest whereas the norms were developed on students taking it for the first time. This is one of the many problems in using test norms for program evaluation to be discussed under Issue 6.

The more practical reason for using cross-sectional designs is to shorten the data collection period. While there is no problem in commissioning evaluations that measure pretreatment achievement levels in the fall and posttreatment achievement levels in the following spring, it is much more costly, in many ways, to measure pre-to-post gains over a period of several years to test rationales based on long-term effects of compensatory education.

The primary issue of how long after a Title I treatment one should measure the effects of that treatment has proven to be an important issue, because of reported results (e.g., Thomas and Pélavin, 1976; Pelavin and David, 1977) that students show good progress when measured from a pretest in the fall to the posttest in the spring of the same year, but looking over the longer trend, the students who are Title I participants tend to fall further and further behind with each grade. For this reason, the question of whether there are long-term, sustained effects of Title I is not answered by evaluations of short-term gains. The evaluation of these effects is the goal of the current evaluation of the sustaining effects of compensatory education being carried for USOE by System Development Corporation.

The question of whether Title I should be evaluated in terms of achievement gains within the school year or over a longer period depends on fundamental systemic questions about the aims of Title I. These aims are not to provide a separate school track for the educationally disadvantaged, in which each grade teaches one set of materials to compensatory students and another to noncompensatory students, but to teach the skills necessary to bring children up to the level of competence necessary to benefit from noncompensatory instruction. The consequences of this view of the purpose of Title I are substantial. For example, it leads to the allocation of funds to the early grades, to ensure that children who start out with a home life that does not provide them with the prerequisites for handling schoolwork successfully will be brought up to a level at which they can cope with their school tasks as soon as possible. The alternative view is that the students who are in Title I will need a continuing special education program because of their lower capacities for learning (for whatever reason), in which case the expectation is that at the end of the first year of a treatment they will have learned more than if they had not had that treatment, but that nevertheless they will need to continue the treatment in succeeding years. In this view, the role of the Title I treatment is not to give them some basic skill that allows them to catch up, but rather to provide a different, more individually adaptive and possibly more expensive curriculum by which they can learn what is needed to be learned at each grade.

At the very least, there must be measurement from one year to the next, because of the various problems stemming from the administering of pretests in the fall and posttests in the spring. There are apparently differential losses of skills over the summer, and it is quite possible that students who learn well during the school year in the Title I program may in fact lose a lot of what they have learned over the summer and come into the next grade further behind their peers than they were at the beginning of the previous grade.

Evidence to support the need for valid, long-term, temporal relational information has come from studies of the long-term effects of Follow Through participants in New Haven, Connecticut (Abelson, Zigler, and DeBlasi, 1974; Seitz, Apfel, and Efron, 1977). In that set of studies, which involved testing children several times starting in kindergarten and continuing through the eighth grade (as of 1977), continued differences between Follow Through and non-Follow Through participants four years after completing the program were shown. Although the samples were quite small and the results not truly dramatic in that study, the use of a longitudinal design did result in demonstration of long-term gains that have not been found in cross-sectional comparisons.

The decision between cross-sectional and true longitudinal designs is complex. Although we can list various advantages and disadvantages of each, the choice in any particular situation will depend on the values assigned to the various advantages at that time.

The primary problem for cross-sectional comparison designs is establishing the equivalence of different cohorts. Among the factors that operate to produce nonequivalence are the following:

1. mobility of students: a significant number of students change schools during the primary grades, and that movement is not random; thus, the sixth graders in a school are not likely to be exactly the same as second graders in that school will become in four years;

2. curriculum variations: the curriculum in a school and the teachers employed who teach the children of the two cohorts will vary; and

3. population trends: a general population trend in achievement scores will confound results.

Another type of problem for cross-sectional designs is that they cannot make use of relations in individual data, but must rely on group means. This means, for one thing, that no information relating variation in personal traits and experiences to variation in long-term gains can be extracted without a true longitudinal design. It also reduces the reliability of the results: the random variation of gains across individuals is substantially smaller than the variation of differences between randomly paired pretest and posttest scores that could be constructed from a cross-sectional comparison.*

A discussion of the use of longitudinal evaluation in educational evaluation has been provided by Ryan (1974). One comparison of results obtained from longitudinal and cross-sectional evaluations of the same educational program (Dyer, Linn, & Patton, 1969) found significant biases in the cross-sectional evaluation methods.

Longitudinal designs take a long time to carry out, however. A compromise option of overlapping panels of longitudinal cohorts is, on the other hand, possibly an acceptable alternative to straight longitudinal designs. As a hypothetical example, over a three-year period three cohorts might be followed: those who in the first year were in grades 2, 3, and 4. In the third year, they would be in grades 4, 5, and 6. Whether this design turns

*The variance of the (longitudinal) gain measure, assuming equal variances (σ_x^2) on the pretest and posttest, is $2\sigma_x^2(1-r^2)$, where r, the correlation between an individual's pretest and posttest score, is likely to be at least .5. The variance of the differences between pretest and posttest scores in a cross-sectional design is $2\sigma_x^2$ (i.e:, $r = 0$) and therefore is substantially higher.

out in fact to be adequate for the particular information need is an empirical question. If on the overlapping periods of such a design there are similar relations among different cohorts (e.g., grades 3 and 4 for the original second and third grades), then it is an adequate design; however, if during the overlapping period there are different relationships, it will be difficult to extrapolate from these overlapping periods to provide temporal relational information between grades 2 and 6. It may, in fact, take up to ten years to perform the correct, valid evaluation of Title I. Keeping this in mind, any contracts for collection of data should be carried out with the assumption that they might be the initial phase of some longitudinal evaluation that would be completed by some other contract in later years. Thus, for example, identities of particular students should be recorded, although carefully guarded from unintended uses, and periodic efforts to follow the movement of students among schools should be undertaken. This would allow the evaluation of a program at the later years to be done in a reasonable time frame for practical policy-making.

In addition to the problem that they take too long for many decision-making purposes, longitudinal studies also incur the costs of correcting for attrition of various types of participants in the evaluation. First, students may not be available for all testing sessions, and omitting them may seriously affect the findings. Trismen et al. (1975) found, for example, that even within a single school year approximately 10% of the students had either pretest-only data or posttest-only data. The students who had missed one or the other test were not a random sample, for they tended to score lower than the students producing complete data. A method for dealing with this attrition, nonrespondent sampling, will be discussed under Issue 3.

A second type of attrition is among teachers, administrators, and even projects being evaluated. If an evaluation measures performance of a set of traetments over several years, it must "correct for" the fact that the treatment will inevitably change over years. A third type of attrition is of evaluation project staff. To ensure that the project will not be subject to breakdowns if individuals change jobs and are replaced, careful records of events and procedures (such as telephone conversations) must be kept that would not be necessary for a project of short duration in which staff attrition would be unlikely.

Summary. Although it is impossible to give a single answer to the topical question of this issue ("is longitudinal evaluation necessary?"), it is possible to make several recommendations based on the experiences of previous educational evaluations.

1. Individual achievement gains should be measured for intervals of whole years to avoid distorting effects of time-of-year (e.g., differential amounts of experience with the teacher giving the test).

2. Conclusions about pretest-posttest gains should not be based on comparison with published norms, because the latter were obtained on children who took the test only once.

3. Teachers' retrospective judgments of children's gains should be ignored. That does not mean that teachers' observations recorded throughout an evaluation period need be ignored.

4. Longitudinal studies of long duration, making use of overlapping cohorts where possible, are necessary for the ultimate impact evaluation of Title I. Such studies are relatively quite expensive, but whenever the information they provide is needed in valid form, avoiding them is short-sighted.

5. Any evaluations undertaken without funding for long-term longitudinal data collection should nevertheless take fairly inexpensive steps to ensure that the data base acquired can later be used as the first stage in a longitudinal study.

Sampling

Introduction

Gathering information to test decision rationales is costly, and program managers and evaluators should weigh the cost-effectiveness of different information gathering plans much as they would weigh the cost-effectiveness of different program strategies. A crucial step that determines the cost and effectiveness of evaluation is sampling. Sampling refers to the process of selecting a few units from which to gather information (e.g., schools, classrooms, or children) from a large population. There are many variations of sampling, and the choice among them must be cognizant of both the cost components of data collection and the nature of the information needs to be satisfied in order to provide maximally effective use of evaluation resources.

The need for sampling in the evaluation of Title I is apparent when one realizes that information is needed on school districts, schools, and school children in order to formulate policy alternatives. There are over 17,000 school districts in the country, approximately 90,000 schools, and over 40,000,000 school children, of whom over 40% attend schools receiving Title I assistance.

There are two categories of sampling: formal and informal. Formal sampling refers to the process of defining a population (e.g., all second graders in Title I assisted schools) and then prescribing a "sampling rule" that determines which units in the population will be observed. That rule normally contains a "random" process, but may be "systematic". Informal sampling refers to the selection of units to be observed without clear specification of the population and the sampling rule. The advantage of formal sampling is that it provides a basis for evaluating how precisely the information gathered on a sample reflects a population. Although it is customary for policy decisions to be made on the basis of information from informal samples, any support for a rationale based on an informal sample is subject to the criticism that the information gatherer may have deliberately selected units to prove his/her point; such an argument is much weaker when a formal sampling procedure has been specified. An informal sample is sufficient only when generalization to a population is unnecessary; for example, a search for effective projects may appropriately be informal if the objective is merely to find a few, but must be formal if a conclusion is desired concerning the frequency of effective projects in a population.

A formal sampling procedure will yield a probability sample that is representative of a population if the relative frequency (probability) of each unit being selected is known and greater than zero. Among probability sampling methods, there are numerous variations that aim to use information known about the population in order to reduce the cost of obtaining information. The basic method is simple random sampling with replacement. In order to select such a sample, one needs a numbered list of the units in the population and a way of generating a list of (pseudo-)random numbers (e.g., a table in a statistical textbook). Each successive unit is selected for observation if its number appears on the list of random numbers. The statistical computations are simplest for this method of sampling. The first variant is sampling without replacement, in which if a particular random number occurs more than once on the list, the corresponding unit is nevertheless only selected once. Because collecting repeated information on the same unit causes interpretive difficulties, this variant is nearly universally used, although in practice if the sample is less than 5% of the population; the two methods should produce essentially the same conclusions.

There are four more substantive categories of variation in probability sampling: stratification, clustering, multistaging, and proportional sampling. We shall only describe them briefly here; the reader who wishes further information can consult a textbook on sampling (e.g., Hansen, Hurwitz, and Madow, 1953; Cochran, 1963; Raj, 1968). Stratification refers to the use of knowledge about some factor on which the units in the population vary (e.g., region of the country) in order to ensure that exactly the right number of units is selected from each "stratum," or level of the factor. Stratification can serve two purposes: (1) to increase the precision of information gathered by eliminating a portion of the random error, and (2) to allow more frequent sampling from some strata than others in such a way that mathematical corrections maintain the representativeness of the sample. Clustering refers to the sampling of some superordinate units in order to select units to observe. For example, all the major evaluative studies of Title I that have reached conclusions concerning children participating in compensatory education have first selected school districts (USOE, 1970; Glass, 1970; NCES, 1975, 1976; GAO, 1975) or schools (Trismen et al., 1975), and then observed only the children in those selected clusters. If within selected clusters, only a sample of the units of interest is to be observed, then the sampling procedure is called multistage. The purpose of clustering and multistage sampling is to reduce the cost of collectin, data; for example, test administration costs are more closely related to the number of testing sessions required than to the number of children tested in each session, and children in a single classroom can all be tested in a single session. The fourth major variation in probability sampling is cluster sampling with probability proportional to "size." In this variation, the probability of a particular superordinate unit's being selected is proportional to the number of units of interest it contains. For example, selection of school districts might be undertaken based on the average daily membership of the districts, so that a district serving 20,000 students would have 50 times the probability of being selected as a district serving 400 students. The purpose of sampling with probability proportional to "size" is to maximize the precision of information on the population of interest (e.g., students) while minimizing the number of clusters that must be contacted in collecting the data.

All of these variants improve the efficiency of information gathering over the basic method of simple random sampling. The costs associated with them are (1) that they require some further information about the structure of the population to be sampled; and (2) that the interpretation of the data from more complex combinations of the variants is more complex, in some cases beyond the limits of current statistical sophistication.

The first issue to be discussed in this section concerns (1) the relation of information needs to the need for a probability sample and (2) the threats to representativeness that must be dealt with.

The second of the two issue discussed in this section concerns the relationship of cost of data collection to sample size and the relationship of sample size to the precision of the information produced.

Issue 3. When is representative sampling important?

The need to generalize results from a sample to a populatirn depends on the decision rationale being tested. There are at least three distinctly different types of information need that require lifferent levels of representativeness: (a) the need to know the average value or frequency of an event in a population (e.g., the average class size of Title I assisted classrooms); (b) the need to know whether two or more variables are related to each other (pcssibly casually) (e.g., an instructional method and amount of student progress); and (c) the need for some examples of a type of event (e.g., a suc:essful project). In discussing thio issue, we shall consider both the levels of representativeness needed for each type of information and the two major threats to representativeness that must be dealt with: misinterpretation based on confusion of units of analysis and misinterpretation based on lack of usable data provided by some of the selected units (i.g., nonresponse bias).

For the first type of information need, estimates of parameters of program operation, strict quanti'ative representativeness is a necessity. For that reason, the results of the TEMPO study (Mosbaek, 1968), the aggregations of annual state reports on Title I (Wargo et al., 1972; Gamel et al., 1975; Thomas and Pelavin, 1976), and the GAO study (1975) cannot be accepted as quantitatively accurate pictures of national program operation. The USOE surveys (USOE, 1970; Glass, 1970), and the NCES surveys (1975, 1976), on the other hand, do provide quantitatively accurate generalizations to the national population, insofar as the information gathered from the samples was accurate.

Turning to the second type of information need, whenever the conclusions to be reached concern the existence of ations that should appear within any given project, such as between methods and impact, it is not essential that the project(s) observed be quantitatively representative of a population. The conclusions would be questioned, however, if the projects selected were especially unusual on some dimension; therefore, some effort is worthwhile to select a project or projects that are reasonably representative of a population to which one wishes to generalize. Obvious examples are experimental demonstrations tha' are selected for the particular processes to be investigated; the implied goal of such studies is to determine better methods for compensatory education that can be used by the school system at large. As part of the Compensatory Reading Study, a sample of schools that were either especially effective or especially ineffective and that varied across clusters of methods used was selected for in-depth observation. From that investigation, the researchers were able to identify attributes characteristic of effective schools. While the results cannot be guaranteed to generalize to all school districts, they serve a useful purpose in the incremental increase of our general understanding of how to design compensatory education projects. Other research studies, such as M. McLaughlin's (1971) and those cited by Gordon and Koutrelakos (1971), provide quite interesting recommendations for improving compensatory education, and although there are grounds for questioning the validity of their results from a design perspective, the lack of a representative national sample is not one of these grounds.

As an example of a hypothetical case in which achieving quantitative representativeness could actually distort the results of a relational study, consider the data in Table 4. If two factors, A and B, are correlated in the population, then reflecting that correlation in the sample, as shown by the columns labeled "N" in the table, could result in a spurious conclusion, in this case that Factor B was a predictor of scores ($\bar{X}$). Examination of Table 4 shows that Factor B is not truly directly predictive of scores; only through its association with Factor A is it correlated with scores. Although data collected according to representative sampling rules can be treated statistically to produce undistorted results concerning relations, that treatment can be quite complex. Data collected according to nonrepresentative but orthogonal (uncorrelated) sampling rules are easier to interpret.

Table 4

Hypothetical Example of a Distortion
Produced by Quantitative Representativeness

		Factor A				
		Low		High		Total
		$\bar{X}$	N	$\bar{X}$	N	$\bar{X}$
Factor B	Low	10	160	20	40	12
	High	10	40	20	160	18

The third type of study, popular in the federal educational administration because of its potential for producing large benefits, is the search for successful, exemplary projects that can be packaged and disseminated. Representative sampling is not needed to satisfy this type of information need. It is much more efficient to use any informal sampling methods available, such as consulting prograr experts, in order to focus observation on the successful projects. This type of study has had a recurrent problem, however, that may be due either to problems with the method of identifying outstanding projects or the problem of capitalizing on chance occurrences: later evaluations have in many cases not clearly corroborated the success of the projects identified earlier as exemplary (Wargo et al., 1971; Stearns, 1977).

To summarize the needs for representativeness, the method of selecting a sample for an evaluation study is dependent upon the objectives. Studies aiming to identify relations among processes and outcomes should avoid random, representative sampling in favor of sampling for significant variation in processes and outcomes. Studies aiming to assess parameters of program operation statewide or nationwide, on the other hand, must obtain representative samples in order to provide accurate, unbiased information. For example, we would not require a study that found individualized instruction to produce reading gains to have a nationally representative sample, but we would require representativeness of a study that reported that blacks tended to receive compensatory instruction relatively more frequently than whites. In general, this issue is not as controversial as some others, primarily because the methodological problems have apparently been at least approximately solved.

Turning now to the threats to representativeness, the first threat (misinterpretation based on confusion of units of analysis) is a semantic problem that merely requires sophistication on the part of the evaluator to avoid erroneous statements of conclusions. The second threat (misinterpretation based on nonresponse bias) is a substantive problem requiring careful attention in the planning and execution of data collection as well as careful interpretatiomof data.

The problem of confusion of units of analysis arises when one uses clustering or multistage sampling. The simplest way of avoiding confusion is to state results in terms of an "observational" unit that is equivalent to the clustering unit. Observational units are units that are referred to in statements summarizing the results of the evaluation. Thus, a statement in the conclusion of an evaluation report might be either "compensatory education projects in the sample tended to vary greatly in ..." or "compensatory education students in the sample tended to vary greatly in ..." Each statement presumes a particular type of observational unit. Sampling units, as opposed to observational units, are the units whose relationship to a population of interest is known. It is important to establish the observational unit that is crucial for the information needed and then to select sampling units so that statements can be validly made in terms of thos observational units. The possible observational and sampling units for Title I include:

1. students,

2. teachers,

3. groups of students receiving a particular service,

4. classrooms,

5. schools,

6. school districts, and

7. states.

Is it reasonable to sample schools within a state and make statements about students? The answer is generally "yes." However, when the schools do not exactly represent the proportions of students in the population for which generalizations are to be made, then the mean scores for the schools must be weighted differentially during aggregation.

Basically, if the observational unit is to be students, then to produce stable, unbiased estimates for the population of students based on a sample of schools (and testing of a specified set of students in each school), it is most efficient to select schools in such a way that the likelihood of each school being selected is proportional to the number of students in the school.

A problem that can arise if one is not careful in using differing observational and sampling units (e.g., students and schools) is in making observational statements that in fact depend on the way in which observational units are distributed within sampling units. Such an error occurred in the Compensatory Reading Study (Trismen et al., 1975). The authors noted (p. 75) that minority disadvantaged students tended to receive compensatory instruction in separate classrooms, while white disadvantaged students tended to receive it in classrooms combined with non-disadvantaged students. But their conclusion, "it seems that such student assignments are being made at least in part on the basis of ethnicity," overlooks the structure of their sampling. In fact, it is equally plausible that these effects were between schools and that schools with especially large minority enrollments were also those that, for other reasons, had chosen to use separate rather than combined classes for compensatory reading instruction. This possibility would have been easily testable had the analyses taken into account the difference between the sampling method (by schools) and the units about which the statement was intended to be made (students).

To summarize, clustering or multistage sampling requires some care in interpretation of data that is not necessary in studies employing simple or stratified random sampling. Otherwise, conclusions can be reached and rationales supported that are in error.

The other threat to representativeness is nonresponse. This important aspect of sampling, which occurs in practice but is not usually covered in elementary statistical texts, is the problem posed by sampled units that do not choose to participate. For example, in the Compensatory Reading Study, 731 school districts were carefully selected (in Phase I) as candidates for the sample, but then only the first 222 who responded that they were ready to be involved in the study were actually included (in Phase II). Another way in which nonresponse bias can occur is through the reporting of invalid or unusable data. The summaries of annual state reports (Wargo et al., 1972; Gamel et al., 1975; Thomas and Pelavin, 1976) have suffered from the fact that although reports were available for the large majority of states, most of the reports did not provide the quantitative information needed to produce a national summary, especially of achievement gains from Title I. That nonresponse bias can be important for some variables and not others was shown in the national Title I surveys of 1967-68 and 1968-69. In these surveys, although response was good for questions of participation and service delivery, it was completely inadequate for questions of impact on achievement -- only 6% or 7% of the districts provided adequate achievement results.

This kind of sampling problem, nonresponse bias, is difficult but not impossible to handle. The first step is to compare what data are available from the nonresponding units to corresponding data on responding units to test whether responding and nonresponding units are really from different populations. If no difference is found on a variety of characteristics related to the variables of interest, nonresponse may not contribute a great deal of bias to the study's results. Also, if nonresponse is limited to fewer than 10% of the sampled units, as a rule of thumb, then the bias introduced is likely to be unimportant.

Some differences between units that do and do not respond are very likely to be observed, and a nonresponse rate of greater than 10% is frequent. There are two solutions in this case. (1) If on various stratifications of the sample there are at least some units in each cell who respond, then the results from the units that respond can be weighted accordingly to stand for both themselves and the units that did not. For example, if in stratum A of a sample of schools, 4 of 10 schools participate, and in stratum B, 8 of 10 participate, each score in stratum A should be weighted by twice as much (the ratio of 8/10 to 4/10) as the scores for schools in stratum B. (2) One can choose a small sample of the nonparticipants and by intense efforts gain their participation. From these comparisons, estimates of nonresponse bias can be obtained. Such nonrespondent sampling and follow-up is crucial to the validity of any estimates of population statistics when fewer than 75% of the sampled units agree to participate and do in fact produce usable data.

Nonresponse bias is especially a problem for longitudinal studies. When gains are measured from pretest to posttest, the mobility of children between schools can substantially affect the conclusions reached -- if children who leave a particular sampled school tend to learn more slowly than those who remain, apparent gains will be greater than if all the children were tested at both times. It is clear that in order to provide meaningful analyses of pretest to posttest gains, the same students must be included in both pretest and posttest samples. This means, based on the examination of nonresponse bias in the Compensatory Reading Study (Trismen et al., 1975), that the children included in such analyses will tend to be substantially less educationally disadvantaged than the totality of children participating in compensatory education. In the Compensatory Reading Study, the choice was made to analyze gains for instructional group means that included all children who took either the pretest or posttest. Although that choice ensured that the most disadvantaged children were included in the analysis, the meaningfulness of "gains" computed between pretest and posttest groups containing different children is highly questionable: any gains would be confounded by mobility effects.* The only apparent solution to the mobility problem is to analyze the data according to a more sophisticated model that treats student mobility and other causes of nonresponse as components of the system and evaluates them as well as achievement gains of students who take both pretests and posttests. This would require tracking down and posttesting at least a small representative sample of pretested students who are not present for the posttest.

There are two general recommendations that follow from the points made in the discussion of this issue. From these, many specific recommendations for procedures can be derived.

1. Sampling plans for evaluation should be carefully related to the information needs to be satisfied. Nationally representative samples are necessary only when quantitative estimates of program operating characteristics are needed, and they may impede the gathering of certain other types of information.

2. Plans for the analysis of data should be carefully examined prior to sampling for their implications on sampling procedures, and vice versa, so that the problems associated with use of different observational and sampling units and with nonresponse bias can be foreseen and dealt with in the context of a single comprehensive system model. Only then can the data collected by comfortably accepted as representative of program operation. This recommendation goes beyond sampling and will be elaborated in the discussion of measurement and analysis issues.

* The use of instructional group means in the Compensatory Reading Study also suffered from the fact that the few children who were in compensatory classes in the fall and regular reading classes in the spring (presumably because they improved significantly) would have their pretests counted in the compensatory group means and their posttests counted in the regular group means.

Issue 4. How large a sample is necessary?

The choice of sample size for federal social program evaluations is largely arbitrary. One can obtain useful information from observing one school or ten thousand. Although a quantitative methodology exists for determining the sample size needed for an evaluation as a function of the precision of the information needed, the need for (and, therefore value of) precision is nearly impossible to quantify. For example, for most policy-making, it is immaterial whether finding that an event occurs 30% of the time in a sample means that 19 times (samples) out of 20 the population percentage would be between 25% and 35% or between 20% and 40%. Yet the sample size would have to be roughly four times as large in the former case as in the latter.

In the case of compensatory education evaluations involving achievement gains, a plausible criterion for information precision has been suggested: that observing a gain which is "educationally significant" in a sample should allow one to infer that at least 19 times out of 20 that gain would not be purely by chance. This criterion depends, of course, on an acceptable definition of educational significance. Brief discussion of the use of this criterion to determine sample size and of the relationship between sample size and information gathering costs is as far as the present consideration of sample size will extend. Readers who wish further information are urged to consult a text on survey sampling (e.g., Raj, 1968).

To determine sample size, we need to consider not only the total sample but also the size of the groups that we want to compare. As the evaluators of Head Start found out nearly a decade ago, it was not sufficient just to obtain a sample of 100 Head Start programs, because it turned out that the sample included only 30 full-year programs, as opposed to summer programs. This did not provide a sufficient data base for statements describing the effectiveness of the full-year programs. If the design of the program evaluation calls for sampling in ten different categories (e.g., grades, project treatment types), the sample size in each of these categories should be determined so that a stable mean can be estimated for that category.

Some authors have proposed that for educational program evaluation a gain or difference of one-third of a population standard deviation be considered educationally significant (e.g., Horst, Tallmadge, and Wood, 1975). While nobody claims that this criterion of educational significance is "correct," the fact that it has been referred to repeatedly demonstrates the need for some such criterion, and research to establish a criterion is called for. For the purposes of sample size determination, we can use this criterion in calculations exemplified by the following simple experimental design. Let us assume that we want one-third of a standard deviation difference between two groups (a Title I treatment and a control group) to be significant at the .01 level on a two-tailed test. That is, we want the likelihood of observing that difference (or larger) by chance alone to be less than one in a hundred. This leads by simple algebra and an assumption about the randomness of the chance effects to an estimate of the sample size.

$$\frac{\left[\text{minimum difference to be detectable}\right]}{\left[\text{within group standard deviation}\right]} \times \sqrt{\frac{\left[\text{necessary size for each group} \; -1\right]}{2}} = \left[\begin{array}{l}\text{critical value corresponding to reliability}\\ \text{of detection desired, from tables of the t}\\ \text{distribution}\end{array}\right];$$

or reordering this equation:

$$\left[\text{necessary size for each group}\right] = 1 + 2\left(\frac{\left[\text{critical value of t}\right] \times \left[\text{within group standard deviation}\right]}{\left[\text{minimum difference to be detectable}\right]}\right)^2 .$$

The critical value of t corresponding to a .01 significance level is 2.58; so, if it is necessary to attribute a difference that is K times as large as the typical random variation of scores within groups, the necessary sample size is given by:

$$N = 1 + 13.3/K^2.$$

If the minimum detectable difference were to be one third of a population standard deviation (determined from published test norm tables) and the standard deviation within each of the two groups being compared were one-half the population standard deviation ($K = 1/3 \div 1/2 = 2/3$), then the required sample size would be 31 in the treatment and 31 in the comparison group. If we were satisfied with a .05 level of significance, the necessary sample size would be about 20 in each group. Thus, it is usually unreasonable to expect that a teacher should be able to evaluate the effectiveness of a compensatory reading program on the basis of his or her students in a single class, because the class will not be large enough to allow detection of some educationally significant differences between treatment and comparison students. Moreover, if that teacher can clearly see such a gain, it must be quite a bit in excess of the minimum needed to be evidence of "eduational significance."

On the other hand, in school districts of moderate size or larger, there certainly would be enough students to be able to carry out an evaluation of their compensatory education project using the one-third standard deviation criterion of educational significance.

We should remind the reader that the selection of the minimum effect to be detectable was arbitrary and it was crucial for the calculation. Thus, it is crucial for the final resolution of the sample size question to determine the exact form of the comparison to be made. To take a different type of comparison, suppose we wished to compare two different treatment groups on the percentage of participants achieving a particular minimum proficiency level. If we wished to be able to reliably (at the .05 level) detect any differences in percentage of 20% or more (e.g., 50% vs. 30% or 90% vs. 70%), an estimate of the required sample size can be obtained as:

$$\begin{bmatrix}\text{percent difference}\\ \text{to be detectable}\end{bmatrix} \Big/ \begin{bmatrix}\text{standard deviation}\\ \text{of the difference}\end{bmatrix} = \begin{bmatrix}\text{normal deviate corres-}\\ \text{ponding to .05 level}\\ \text{of significance}\end{bmatrix}$$

$$\begin{bmatrix}\text{percent difference}\\ \text{to be detectable}\end{bmatrix} \times \sqrt{2\begin{bmatrix}\text{size required of}\\ \text{each group}\end{bmatrix}} = 1.96;$$

$$\text{or } N = \frac{1}{2}\left(\frac{1.96}{.20}\right)^2 = 48.$$

Any calculations of sample size are critically dependent on the needed minimum level of reliably detectable effect. In tradeoffs with other cost dimensions, evaluation designers should decide with program managers what precision is needed in terms of the use to which the results are to be put.

There is another aspect of sample size that must be considered. Any evaluation of a program such as Title I is carried out over a particular geographic and demographic area. A school district may be interested, for example, in evaluation of the program within its district, a state within its state, and the USOE and Congress may be concerned with evaluation across the whole country. In each case, it is not sufficient to sample a single unit, such as a school, even though there may be a sufficiently large number of students in that school, because the particular attributes of that school might be quite different from the attributes of schools across the district, the state, or country; these differences might lead to quite different conclusions with respect to the effectiveness of Title I, depending on which school was chosen. Thus, the sample must include units chosen to represent the total variability across

$$
\begin{aligned}
\text{Cost} = {} & \$A_1\ (N_1)\ (L_1)\ (H_1) + \ldots + \$A_K\ (N_K)\ (L_K)\ (H_K)\ , \\
& + \$B_1\ (N_1)\ (L_1) + \ldots + \$B_K\ (N_K)\ (L_K) \\
& + \$C_1\ (N_1)\ (L_1 - 1) + \ldots + \$C_K\ (N_K)\ (L_K - 1) \\
& + \$D_1\ (H_1) + \ldots + \$D_K\ (H_K) \\
& + \$E
\end{aligned}
$$

Notation:

The subscript 1, ..., K refers to different classes of individuals who must be contacted or tested during data collection, such as students, teachers, local school administrators, and state administrators.

N refers to the number of each type of individual involved;

L refers to the number of contacts over time with each individual; and

H refers to the depth, or length of each contact.

The costs are:

$\$A_i$ is the cost per unit time (or depth) of collecting data from individuals of type i, once one has contacted the individuals;

$\$B_i$ is the cost of each locating and reaching an individual of type i;

$\$C_i$ is the cost of keeping track of him/her for subsequent data collection, in a longitudinal design;

$\$D_i$ is the cost of preparing the contact and data gathering procedure for individuals of type i; that is, the instrumentation cost; and

\$E is planning, management, analysis, and reporting cost.

Figure 4. A first approximation to estimation of information-gathering costs in an evaluation.

the population, and the necessary sample size would apply to the number of schools selected, not the total number of students tested. This implies that costs are not merely for testing each student, but rather that costs associated with setting up observations at each school or district, irrespective of the number of students tested, must be included.

Having established the size needed for a study, the cost of it can roughly be estimated using a computation of the form shown in Figure 4. Clearly that figure is an oversimplification, which can be refined dramatically for different types of evaluation. Comparison of the cost with an estimate of the benefits to be gained from the evaluation would provide a rational method for deciding whether to carry out the evaluation. On the other hand, in the real world in which the benefits from evaluation are nearly impossible to estimate beforehand, the comparison is usually with a prespecified budget allocation for evaluation. In the case in which the estimated cost exceeds the budget allocation, which is the most frequent situation (at least from the point of view of proponents of planning and objective, rational decisionmaking), decisions must be made of which information needs should remain unfulfilled in the study or what precision should be sacrificed.

In summary, the point of this discussion is first to demonstrate that there are methods for determining sample size from knowledge of information precision needs and information costs, but second, to note that the specification of information precision needs is still only vaguely understood in educational evaluation.

Measurement

Introduction

Measurement refers to the process of assigning numbers to represent constructs, objects, or events of interest. The purpose of assigning numbers is to make it possible to aggregate and compare different events easily (e.g., it is easy to compare two test scores, but can be laborious to compare the unstructured behaviours of two students in a classroom). There is an extensive literature on the mathematical foundations of measurement of which an expert evaluator must be knowledgeable, just as he or she must be owledgeable of the mathematics of experimental design, sampling, and data analysis. The general purposes of that literature are (1) to develop new methods for measurement and (2) to establish and delineate the meaningfulness of conclusions based on measurements. The principle underlying the second purpose is that measurement should not distort reality; conclusions based on comparisons of numbers resulting from measurement should be the same as the conclusions one would reach if the constructs, objects, or events being measured were directly compared without assigning numbers.

The measurement issues to be discussed in this section concern the impact of compensatory education on educational disadvantage. Knowledge of the intricacies of cost and expenditure measurement are also of importance for program evaluation; readers who wish to find out about these intricacies in the context of compensatory education evaluation should read the cost analysis report by Dienemann, Flynn, and Al-Salam, (1974). The problems of testing are the more controversial measurement issues related to compensatory education, however, and will receive major attention here.

Achievement measurement is the central task in the evaluation of compensatory education programs. At a recent national conference on standardized achievement testing of disadvantaged students (Wargo and Green, 1977), Wargo noted that:

> A major reason for the increased use of standardized achievement tests in elementary and secondary education program evaluation relates to the general thrust of school aid at those levels. Most federal financial support programs for local educational agencies have as one of the primary objectives, if not their primary objective, the overcoming of educational disadvantages suffered by students from low socioeconomic status

> families or from culturally different backgrounds. The translation of such legislative goals into program objectives usually means a focus on improving the basic skills (reading, writing, and mathematics) of such students. That combination of legislative and programmatic thrust serves as a major impetus for evaluation specialists to select off-the-shelf standardized achievement tests for determining local, statewide, and national education program impact. (p. 4)

Also, the U.S. Office of Education's current major efforts to provide technical assistance to states and local districts in their Title I evaluations centers around a set of models for collecting and analyzing achievement data.

Deficiencies in the measurement of achievement have shared with deficiencies in use of control groups (Issue 1) the major focus of controversy surrounding evaluations of compensatory education. Other measurement issues in Title I evaluation do not meet the political stress engendered by the fact that certain ethnic groups tend to score lower on achievement tests than others. Furthermore, because achievement tests are frequently used as mechanisms of personnel selection for high-paying jobs and higher education, there is an implicit threat in the use of achievement tests in program evaluation that the individual's scores will somehow later be used against him/her.

The consideration of measurement issues is divided into three parts. First, there is the problem of identifying and selecting which constructs to measure; should one, for example, measure overall progress in "learning to read" or should one measure component skills learned? Also, to what extent is it the role of evaluators to measure noncognitive benefits and side-effects of program operation? Second, there is the selection of an instrument; although that theoretically should follow after selection of constructs to test, the usual situation in practice is that the availability of tests determines which constructs are tested. A very controversial aspect of the instrumentation issue is whether or not to use criterion-referenced tests. The third issue concerns the manner of recording of scores to be used in analysis. As such, it is on the border between measurement and analysis issues. However, because its controversial aspects relate to the content of test publishers' manuals rather than to experimental design, we have included it in this section. A subtitle for the issue: "Are grade-equivalent scores really that bad?" reflects the focus of controversy on this issue.

The aim, as in earlier sections, is to inform the reader of the content of the issues, to point out the critical problems, and to suggest ways in which the issues may possibly be resolved.

Issue 5. What constructs should be measured to determine Title I impact?

Within schools in low-income areas, Title 1 prescribes that services are to be provided to educationally disadvantaged children in order to "meet their special needs". Educationally disadvantaged children have been defined as those who are judged not to be likely to be graduated from high school (USOE, 1970; Glass, 1970), or who are judged at least a year behind the achievement levels expected of their age group (GAO, 1975), using subjective judgments or scores on achievement tests. Special instructional services are to be provided to all the specified children, and special noninstructional services can be appended to the program to supplement the instructional services. Therefore, measures of impact must reflect the extent to which achievement levels are improved by the program, and the constructs measured must be those that relate to achievement. That does not imply that achievement test scores are the only criterion for impact evaluation. In fact, children are in schools for a dozen years or more, and achievement levels in higher grades may depend on many factors other than achievement in the first few years of school. (1) What factors are related to achievement? (2) Should achievement be measured in wholistic terms (e.g., can Johnny read?) or in terms of component skills? (3) Should achievement be measured in terms of scientific theories of achievement or in empiricist terms of "what achievement tests test?" Until such questions are addressed, impact evaluations will suffer from charges of "narrowness" and "superficiality" and even "irrelevance" of their outcome measures, and therefore of their conclusions. The discussion of this issue will focus on these three questions.

The first question, in practice, concerns the relationship between attitude and achievement. Improving children's attitudes is viewed by many compensatory education teachers as an important objective for their activities -- they believe that its ultimate payoff in terms of achievement may be much greater than the learning of a few specific components of reading. The evidence is mixed concerning that relationship, however. Shavelson, Hubner, and Stanton (1976) cited studies that empirically support the notion that improving a child's self-concept will lead to achievement gains. Project LONGSTEP (Coles and Chalupsky, 1976, Vol. II) found a positive correlation between an attitude composite and achievement scores; however, the Compensatory Reading Study (Trismen et al., 1975) found a negative correlation. The degree of standardization of attitude measures is as yet insufficient to allow one to compare these different results; in any case, before attitude measures can become acceptable, indicators of ultimate achievement effects, a substantial amount of research into the strength of that relationship - and into the ways of enhancing the relationship - is necessary. Thus, our conclusions are (1) that attitude measures can play only a supplementary role to achievement tests at present for determining whether a Title I treatment is having impact on achievement, but (2) that it is likely, when adequate research is available, that some kinds of attitude improvement will be shown to be a reasonable short-term goal for treatments that aim for long-term achievement gains, so attitude measurement should not be discouraged.

Assessments of achievement in Title I evaluations have tended to focus on reading, language arts and mathematics. The question of whether it is achievement in general or the mastery of particular skills related to achievement that should be assessed in these evaluations is of concern to specialists in each of these areas. In order to simplify discussion, we have (like Trismen et al., 1975; GAO, 1975; and Thomas and Pelavin, 1976) chosen reading achievement as the example from which generalizations can be made to language arts and mathematics achievement. The second issue referred to above is whether or not it is reasonable to assess reading achievement in terms of specific skills (e.g. decoding, memory, inference, visual acuity, specific vocabulary), each of which alone does not constitute the ability to read, but that are component skills that are believed to contribute to reading achievement. The case has frequently been made (e.g., Stearns, 1977) that standardized tests such as the Metropolitan Achievement Test and the California Test of Basic Skills almost completely fail to capture the content of particular remedial or compensatory reading programs. The reason given for this failure is that Title I teachers typically focus their efforts on specific skills that are related to reading achievement rather than on reading achievement itself. If the participating children have clear needs for which such intense focused effort is warranted, which is undoubtedly the case for many, then assessment of progress in terms of tests most of whose items require skills not addressed by the treatment seems unfair. On the other hand, focusing on a particular component skill may not ultimately enhance reading achievement. As with attitude outcomes, it seems necessary to include in the evaluation of a treatment some measure of overall reading achievement (possibly one or more years after the treatment, which is not in conflict with the need for annual evaluations).

The third question to be addressed in this discussion is whether evaluations should be firmly based in scientific theories of (reading) achievement or whether they should be firmly based in empirical pragmatism: measuring what test publishers call achievement. Of course, firm grounding in theory is preferable - if the theory is correct. There are many theories, or models, of the process of learning to read, however, and at least some of them must be wrong. In fact, it is likely that there are many different ways to learn to read, even for a single individual, so measurement would have to be in terms of alternative theories for learning to read. Williams (1973) has reviewed models for learning to read and lists six categories of theories: taxonomic, psychometric, behavioral, cognitive, information processing, and linguistic. A synthesis of the many perspectives on cognitive achievement is clearly needed as an initial step, if we are to be able to evaluate impact directly in terms of the achievement of new cognitive skills rather than indirectly in terms of the possible use of those cognitive skills to answer questions on an "achievement test". It should be pointed out, in fairness to the developers of commercial tests, that many of them have, especially in recent times, attempted to select items for tests in such a way that scores for particular subscales of items can be interpreted in terms of specific skill mastery.

The value of a firm grounding of compensatory education evaluation in the theory of cognitive achievement should be clear. Such controversies as to whether students participating in a Title I treatment should be expected to learn 70% as much as the median student in a particular time period, or 90% or 110%, are based on a lack of knowledge of just what types of skills should be learned and are being learned by individuals who at the beginning of treatment have some other particular set of skills. In terms of an adequate theory, an individual child's level of achievement could be characterized either as the constellation of skills that he or she has acquired, or for the purpose of summarization, the proportion he or she has completed of the total learning effort needed to reach an ultimate achievement goal. Although the research needed in order to implement this approach is quite substantial, it would appear to involve no scientific procedures that are not presently feasible.

To summarize our conclusions concerning the selection of contructs to measure in evaluating Title I impact, (1) it appears reasonable to use attitude and other noncognitive measures as supplements to achievement measures, although substantial further research on the relationship between cognitive and noncognitive measures is needed; (2) the same conclusion holds for component skill measures as for attitude measures — they should be supplements to overall achievement measures; and (3) evaluation will be much more useful when based on a scientific theory of cognitive achievement; however, the research to develop a sufficient theoretical framework is substantial. All three of these conclusions are similar in their ambivalence; what we have now is minimally adequate, but with some research into the processes that Title I is intended to affect, a significant improvement in impact evaluation would be possible. Until that research is undertaken, skeptics of evaluation will have reasonable arguments that the use of any particular measurement instrument yields results that too narrowly define the purpose of Title I, or that are irrelevant to the goals of particular Title I treatments, or that are too superficial to capture the essential impact of a treatment.

Issue 6. What types of achievement measurement instruments should be used in Title I evaluation?

There appears to be no reasonable and efficient alternative for measuring program impact on a student's achievement level to requiring him/her to produce answers on a paper-and-pencil test. There are literally thousands of alternative tests, and any teacher may construct a new test to fit any occasion. The major alternatives for test selection are (1) between a locally developed test and a standardized test and (2) between a criterion-referenced test and a non-criterion referenced test. The choice must be made in terms of the particular objectives of the evaluation and will reflect a tradeoff of some values for others. For the choice between a locally developed and a nationally standardized test, the relevant factors are: (1) the credibility inherent in use of a test being used by many others, (2) the availability of norm distribution tables for the standardized test, (3) the possibility of tailoring a locally developed test to reflect local objectives and instructional methods, (4) ease of aggregation of data across sites when ~~standardized tests~~ are used, and (5) the relative costs of buying a test from a commercial publisher and generating items locally. For small, informal evaluations, the choices will clearly be different from the choices for a national evaluation whose validity is likely to come under attack.

To choose between criterion-referenced tests and tests not so designed is a matter of some controversy, primarily because of the strong arguments and large investments on both sides. Basically, a criterion-referenced test is one "that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (Glaser & Nitko, 1971, p. 653) or one whose score "has some sort of meaning in itself, irrespective of the scores for specified groups" (Shaycoft, 1976). Items on criterion-referenced tests are systematically derived from a set of objectives or rationales to be measured rather than by statistical item analysis of a large item-pool. Until quite recently, commercial tests were not developed to be criterion-referenced.* Instead, to provide meaning to raw scores, tables were provided showing what percentage of the population achieved each raw score level; that is, the tests were norm-referenced. Note that the concepts of

* That is not to say that good commercial norm-referenced tests have not been designed to contain items whose rationales are that right answers to them indicate the achievement of particular skills (see, for example, Flanagan, 1951).

criterion-reference and norm-reference are not per se incompatible (test scores can have both absolute and relative interpretations); however, the methods of dev..oping the tests are quite different. Norm-referenced tests are developed to be sensitive to individual differences among students, whereas criterion-referenced tests are developed to be sensitive to degrees of skill attainment for each individual.

The relevant factors for choosi..g between standardized tests that are norm-referenced or criterion-referenced are: (1) the relevance of the content of the test, of whichever type, to achievement constructs being meas..ed; (2) the type of evaluation comparison being made (see Issue 1); (3) the volume of data desired; and (4) cost and availability. For the informal local evaluation (e.g., weekly progress quiz), a teacher is well advised to emulate the principles of criterion-referenced test development rather than deliberately selecting items likely to demonstrate different levels of achievement among students. The choice for large-scale evaluations is more difficult.

In order to clarify the selection problem, we shall consider various arguments for and against, first, norm-referenced 1d then criterion-referenced tests.

<u>Norm-referenced tests</u> are set of items, the distribution of responses to which is known for a sample representative of some population. They offer both the advantage of enabling test scores t. be interpreted in terms of comparisons tc the population and the advantage of credibility, in that they were not developed by the individual who teaches the knowledge and skills. The criticisms of norm-referenced tests deal almost exclusively either with the appropriateness of the noıming process or with the method of selection of item contents to include in the test. The norming problems may be solvable with sufficient funds, because they stem from incompleteness of the data on which norm tables are based; however, the problems with item selection suggest the need for new kinds of tests.

There are eight specific categories of problems with norm-referenced tests -- they do not necessarily all apply to all norm-referenced tests, but they do apply to many. After listing the eight we shall discuss them in detail.

1. Norms are based on a population different form that for which compensatory education is intended.

2. Norms are not longitudinal, so norms for gains are not directly attainable.

3. Norms exist for only one or two testing dates per grade.

4. Articulation of scores between levels is not well validated.

5. Performance is not criterion-referenced for component skills (Although major publishers are moving to accommodate this need).

6. Items are developed to discriminate among individuals, not programs.

7. Items are developed primarily to discriminate performance levels of the majority of typical children, so the items may not be as sensitive to the patterns of learning of educationally disadvantaged children.

8. Test scores have a smaller error component near the ceiling than near the floor of performance on each form.

The first four problems obviously could be solved by extension of the norming process. Are they important, however? The following are some of the distortions of results that have been suggested to result from these problems. The first problem is that the particular sample being tested in a compensatory education evaluation is not the same as a distribution of children in the norm population with the same scores. For example, in the norm population, extremely low scores may be indicative of some permanent or transient learning disabilities that are predictive of certain learning paths, whereas those low scores in ghetto schools may be the result of environmental pressures. Even though some Title I participants will have been included in the norm groups, they will be a minority of the low scorers because of the economic criteria for Title I funding. Thus, for example, among students at the 20th percentile at the beginning of third grade, those that are likely to be selected for compensatory education treatments (e.g., from low economic status families) may be those that by the end of third grade tend to move toward the 15th percentile while others move upward (or vice versa).

This leads to the second need, for longitudinal norms. This need is clear when we consider that students are geographically mobile, as well as dropping out at the upper grades. Thus, norming must take into account student mobility, or else the achievement of the population will appear to be different from (usually greater than) its actual value. More important, perhaps, is the fact that in a pretest-posttest evaluation design, children taking the posttest will have had prior experience (the pretest) on another form of the test, which experience the members of the norm group lacked.

The third problem, that norming is only carried out for one or two dates in a school year, makes it difficult to measure the effectiveness of treatments over intervals other than between appropriate testing datas. One solution used in practice is linear interpolation or extrapolation: if, for example, the norms are for a seven-month interval, but the pretest and posttest are given six months apart, scores are transformed to grade-equivalents (that is, to a grade level for which the score would be the median) and then multiplied by 7/6, to estimate what the gains would have been for seven months so that the scores can be compared with other treatments.

A second solution, provided by some test publishers, consists of growth curves obtained by curve-fitting procedures. The curves can be graphically used to interpolate or extrapolate gains, assuming the validity of the curve-fitting process.

The fourth problem, articulation of levels, arises because norm-referenced tests come with multiple levels, each designed for a particular range of grade levels. For many evaluations, it may be necessary to employ different levels for pretest and posttest to avoid floor or ceiling effects. To estimate the gain between pretest and posttest, it is necessary to convert the pretest and posttest scores to a common scale for comparison. Tables for that conversion are normally provided by test publishers; however, the empirical basis for arriving at the tables is usually limited. For example, a raw score of 50 on level A may correspond to a raw score of 20 on level B for a sample of beginning fourth graders, but that does not imply that the same conversion would be accurate for students at the end of fourth grade: skills learned in fourth grade (in a particular school) might be more related to items on level A than on level B, or vice versa.

The other four problems, relating to item selection, are more serious. First, because the performance measured on norm-referenced tests tends to involve unspecified combinations of many component skills, these tests are not sensitive to the achievement of specific criteria. Thus, programs of instruction that focus on a small set of component skills are unfairly judged using these tests. This was discussed under Issue # 5.

Another problem is that standard achievement tests have been developed to discriminate among individuals in such a way as to be predictive over the future of the individual. That means that they are developed not to be sensitive to particular variations in curriculum. The main criterion for selecting an item from a pool of reasonable items to include in a test has been its correlatior with the total score, not its correlation with an external (validation) measure of skill attainment.

The next problem is that item development has usually included administering items to a sample representative of the population and selecting those that discriminate best, and as a result, items that are particularly sensitive to the achievement of minority populations bur are not as sensitive to achievement in the majority population have been deleted on item analysis, because they account for too little variance. Test publishers have recently given specific attention to this problem, and it may become less important in the future.

The last problem, which concerns t sts consisting of multiple choice items where guessing is permitted (and how could one prohibit it?), is that the reliability of test scores is greater for scores in the top portion of the distribution for any form. At the low end, guessing accounts for a large part of the variance, while at the high end it accounts for little. This means, among other things, that small gains will be harder to detect in the lower region of the distribution than in the upper region. One sidelight on this situation is that an attempt to use out-of-range testing can appear to have an effect by itself: if disadvantaged 10th graders are given a test for 10th graders and score at the chance level they might appear to be three years behind; if given the form of the test designed for 9th graders, as more appropriate, they might . so score near the chance level, so that their scores would appear to be three years behind the 9th graders, or four years behind their actual grade level. Thus, changing forms can increase (or decrease) the apparent deficit of a student by a year or more. A solution to this problem, for the evaluator, is to select a test on which each student will score in the mid-range. To do this for typically heterogeneous groups of students would require a test made up of several articulated levels and administration that required flexible starting points for individuals of grossly different achievement levels.

The problems of measurement via norm-referenced tests are most serious when the tests are used for relative comparisons between a treatment and a nonrandom, unmatched comparison group or between a treatment group and a norm population. For absolute comparisons and for relative comparisons between randomly assigned treatment and control groups, the problems are not so serious. The reason, in the latter case, is that relative comparison in a randomized design does not depend on norms and problems of item selection will apply equally to treatment and control students.

It would seem at this juncture that there is need for some test development activity targeted at the needs of program evaluation. Because this is expensive, the private sector of the test development system will probably be very inquisitive about the market for such evaluative uses of tests in their plans for test develcnment.

Criterion-referenced tests are sets of items clustered around sets of objectives, or component skills, whose mastery is supposed to be equivalent to correct item responses. In the ideal case, items are selected on the basis that they discriminate perfectly between groups of students possessing a skill and groups not possessing it. In cases of skills involving incremental mastery of a large domain, such as vocabulary, measurement of the objective may be more complex than merely mastery or nonmastery, but may involve, for example, percentage of the domain acquired.

The problems with criterion-referenced tests are primarily in the area of availability and cost. Because the concept has been implemented more recently than norm-referenced tests, fewer criterion-referenced tests of high quality are available. Given this situation, evaluators are tempted to use well-known and long trusted norm-referenced tests. For some forms of evaluation design, such as comparisons with a population standard, the value of criterion-referencing is not readily apparent. In general, however, the arguments for increased use of criterion-referenced tests in evaluation appear fairly strong. In particular, the ability of these tests to detect component skill acquisition addresses the complaint of some teachers (Stearns, 1977) that standardized tests are relatively insensitive to the learning of a few component skills.

Several of the strengths of criterion-referenced tests do carry along corresponding problems, when viewed from a critical perspective, as in the presentation by Kosecoff and Fink (1976). For example, to be fair in evaluation of a program, the correct objectives to be tested must be specified by the teacher, and error in matching tested objectives to instructional objectives will diminish the test's sensitivity to the treatment. Thus, an evaluation will be biased by the teacher's degree of ability to match objectives. As another example, because different treatments have different objectives, aggregation of scores is more difficult than when a single total score is obtained. If different treatments have different objectives, then comparisons of the treatments on a criterion-referenced basis would have to be a two-stage process: comparison of the extent to which each treatment met its objective and also comparison between the objectives. A treatment that failed to meet stringent objectives might be superior to one that succeeded in meeting easy objectives. Third, criterion-referenced testing "would generate information about an enormous number of objectives, thus complicating the management, analysis, and reporting of data" (Kosecoff & Fink, 1976, p. 2-35). The production of too much information during an evaluation is a questionable basis for criticism; given modern computer methods for data management and analysis, the added complexity, which corresponds to the greatest strength of criterion-referenced tests, their sensitivity, would be welcomed by many users of evaluation results.

In conclusion, the selection of an instrument for measuring achievement in evaluations of Title I is dependent on the particular information needs to be satisfied and the constructs selected for measurement. Nationally standardized (norm-referenced) tests have the advantage of greater credibility than locally developed tests, but they have the two disadvantages of (1) encouraging evaluation in terms of comparison of local performance against inappropriate norms and (2) measuring program performance in terms of tests designed to assess overall individual differences in achievement and thus insensitive to many dimensions of treatment effects. Criterion-referenced tests have the advantage of producing substantially more detailed and precise information on the performance of each treatment in terms of its own objectives, but they have the disadvantage that, for the purposes of valid aggregation of results across treatments with different objectives, fairly complex interpretations of the results are necessary.

To the extent that major publishers move to compute norms for criterion-referenced tests and to identify particular component skills that subsets of items on their norm-referenced tests assess (as appears to be the case), this distinction becomes less important: one could select a good norm-and-criterion-referenced test and interpret the results to fit the particular information needs.

Issue 7: What units of measurement should be used, or: Are grade-equivalent scores really that bad?

This issue concerns the first step in summarization of results from testirg: should each student's score be entered into analysis as a raw score, or should some transformation of that score be made first? The problem is not one of cost to the evaluator, at least when using transformations for which tables or formulas are available, but rather one of validity versus communicability; the more technically correct units are nct necessarily those that are easiest to understand or directly relevant to decisionmaking. The resolution of this issue clearly must treat validity as fundamental and strive for maximal communicability among the technically correct units. Communicating wrong conclusions very clearly is worse than no communication at all.

One particular unit that has held widespread popularity but whose technical problems have made it notorious is the "grade-equivalent score." In several major evaluation studies (Wargo et al., 1972; Briggs, 1973; Gamel et al., 1975; GAO, 1975; Thomas and Pelavin, 1976), these scores were used because many state or local evaluations were being aggregated, and the units most frequently reported were grade-equivalencts. In most cases, the authors expressed regret of that fact. To deal with this controversy, we shall focus the bulk of our discussion on that unit, pointing out that various of its problems are shared by one or more of its alternatives. This is feasible because, with one or two exceptions, any technical problem with any unit is also a problem for grade-equivalent scores. The strength of grade-equivalent scores lies mainly in the clear meaning they purportedly convey: a student with a grade-equivalent score of, say, 3.5 is apparently at the level of the median student with 5 months instruction in the third grade; if that score were obtained by a student five months through fourth grade, then the student would apparently be one year behind the national norm for his/her classmates.

The seven major alternatives for measurement units are:

1. raw scores: number of items answered correctly;
2. corrected scores: raw scores corrected for guessing so that a score of zero corresponds to pure guessing, as shown below for a test consisting of items each with k possible answers:

$$\text{CORRECTED SCORE} = \text{NUMBER RIGHT} - \frac{\text{NUMBER WRONG}}{k - 1} ;$$

the proper correction for guessing does not count as WRONG those items for which no response is made;

3. whether a skill is mastered: a dichotomous 1 or 0 score indicating whether the student has or has not mastered the skill according to the test;

4. percentiles: percentage of a peer population (national, regional, local, or any other population deemed appropriate for comparison) that would have achieved raw scores lower than the student;

5. grade-equivalents: the number of school years of experience at which the raw score is the median, anchored at 1.0 for the beginning of first grade and altered by attributing one month's schooling to the summer quarter so that there are 10 school months per year to simplify communication; between dates of actual test norm data collection, estimated median scores are obtained by curve-fitting procedures;

6. normalized standard scores or normal curve equivalents: transformation of percentiles to normal deviates (in particular, but not necessarily*, so that the mean score is 50 and so that 99% of the scores are less than 99); and

7. growth scale scores: a transformation of normalized standard scores on different test levels (grade levels) to a common metric, so that a student's growth can be plotted continuously across levels of a test.

No matter which of these measures is used, questions of how to compare pretest and posttest scores or scores between groups remain. These are discussed under Issue 8. We now turn to the specific problems of grade equivalents and their competitors.

It is common to report a student's achievement as equivalent to the median performance of students at a particular grade level. Thus, for example, a student halfway through the fourth grade who was having great difficulty might be described as "a year behind." This is a metric that is apparently

*The term normal curve equivalents was developed by RMC Research Corporation and refers to the specific transformation mentioned. The more general concept is referred to as normalized standard scores.

independent of any test, of any particular curriculum, and of any particular norm group. Moreover, it suggests to a parent the amount of effort needed to bring the student "up to standard." Even though we may criticize the properties of grade-equivalent scores for program evaluation, they serve a distinct purpose for communication of a student's or a class's average achievement in a school year. Thus, test publishers include tables of grade equivalents for the raw scores on their tests. None of the other units have the same clarity and simplicity of meaning, although for two of the units the meaning is fairly direct: percentiles indicate an individual's rank relative to a peer group, and because it is that peer group with whom he/she will be competing throughout life for the best jobs and highest quality of life, "getting behind" and "getting ahead" in percentile terms are meaningful; and indicators of particular skill mastery are directly meaningful to the extent that the skills mastered are directly meaningful (however, some theoretically meaningful skills, such as "decoding" or the Piagetian concept of "conservation," may not be obviously relevant objectives for basic skills instruction for some audiences).

The problems of grade-equivalent scores, as well as other units, stem both from their definition and from their operationalization. The problems stemming from operationalization could presumably be solved with a sufficient expenditure of funds, if the fundamental problems with the concept were not serious. The fundamental problems for grade-equivalent scores derive from the facts (1) that achievement gains are not linear as a function of months in school; (2) that summer period presents special problems; and (3) that the performance of a student a year below grade level is qualitatively different from that of the median student a year younger. The operational problems arise from the fact that norms for standardized tests are published for a single testing time in the school year, or at most two times, so that grade equivalents for most testing dates must be arrived at by interpolation.

The fact that achievement is not linear as a function of time can produce distorted results. In the Thomas and Pelavin study (1976) for example, larger average grade-equivalent gains were reported for compensatory education programs in high school than in the primary grades. Although Thomas and Pelavin did not interpret this effect as meaningful, others might. However, that effect is probably an artifact because, for example, an individual at the 20th percentile might be a half year below grade level in second grade but three years below grade level in tenth grade, so bringing him/her up to the median in a year (unlikely, but taken for simplicity), a gain of 30 percentile points in either case, would show a 1.5 month-per-month gain for the second grader but a 4.0 month-per-month gain for the tenth grader. At another level, learning a specific number of component skills may lead to a 20-percentile gain at one grade level and a 30-percentile gain at another grade level.

The second problem concerns the summer. The lesser problem with the summer is its definition as a single month for the construction of grade equivalents, so that, added to the presumed nine-month school year, it produces a ten-month year in which decimal tenths correspond to months. This clever aid to communication has the unfortunate consequence that grade-equivalents can never be considered quite adequate for use in research on achievement growth patterns because the summer "month" is ill-defined. The more serious problem is that students who are achieving at levels lower than their peers may actually lose ground, in absolute terms, over the summer (that is, they actually have mastery over fewer academic skills at the end of the summer than they had at the beginning of the summer, while the brightest students may gain at a rate at least as great and often surpassing their rate of gains during the school year. (Although this result has not been proven, reports by Kaskowitz and Norwood, 1977, and Pelavin and David, 1977, are highly suggestive.) The result of this difference in students' forgetting and extracurricular learning is to make school-year compensatory education programs seem to have only short-range effects: when measured from fall to the following spring, compensatory education students show strong gains, but the students in the programs year after year may fall further behind their peers. This problem is not merely a problem with grade-equivalent scores but, indeed, with the underlying assumptions of compensatory education, and the issue is discussed further in the synthesis of substantive findings on Title I. However, it causes critical problems for the use of grade-equivalent scores and especially distorts any studies that aggregate results from fall-to-fall (or spring-to-spring) tests with results from fall-to-spring tests.

The third fundamental problem with grade equivalents, and with other scores based on a national norm sample (percentiles, normalized scores, and growth scale scores), concerns the multidimensionality of achievement growth. The assumption implicit in the use of grade equivalents, although not necessary for their construction, is that there is a certain amount to be learned in each grade. In each region of the country and in each classroom, however, particular goals are set that are different, to a greater or lesser extent, from the goals assessed in standardized tests. Among other things, children start school at different ages and have different numbers of school days per year in different states. Furthermore, the amount a fourth grader who is a year behind knows is likely to be qualitatively different from the amount a third grader knows, although their total test scores may be the same. The use of grade-equivalents promotes a simplistic, unidimensional view of achievement. That simplicity must not get in the way of discovery of particular achievements and deficiencies in student and program performance.

A special operational problem for grade-equivalents is that they are based only on data collected at one or two points in the school year. If tests actually are given in an evaluation at other testing times than those for which norming was done, interpolations must be performed to obtain grade-equivalent gains. Thus, if the norms are for September 20 and May 20, eight months apart, and testing is done on October 5 and May 5, seven months apart, evaluators must multiply gains obtained by 8/7 to compare gains occurring in the norm group. The possible distortions caused by such interpolations are so great that test publishers and evaluators have called for all testing to be conducted at the same time in the school year as the norm group was tested. Thus the use of tests with only single norming dates (e.g., in the spring) in evaluations based on fall to spring gains is highly questionable.

The fact that grade-equivalents are based on the performance of average students makes them less useful for studies of students who deviate substantially from the average (e.g., compensatory education participants). It would be preferable to establish expected per-year, or per-month, achievement of students in various percentile ranges, based on longitudinal norming. Then month-for-month gains could be reported for compensatory education students in comparison with students or comparable prior achievement levels.

For raw scores, the fundamental problem is interpretability. The only real meaning for a raw score is its comparison with some other raw score on the same test. If that comparison is the goal of the evaluation, then raw scores may be the most appropriate unit. Raw scores are not guaranteed to have a normal distribution, however, which is required by many procedures; normalized standard scores or normal curve equivalents at least partially solve that problem. (One should note, however, that transforming both pretest and posttest scores to normally distributed scores definitely does not ensure that the resulting bivariate [two-dimensional] scatter plot of scores will conform to the bivariate normal distribution required for some analyses, such as analysis of covariance.)

Correcting raw scores for guessing improves their accuracy by eliminating any biases that might be due to greater tendencies to guess in some groups. Note that this correction for guessing requires that two raw scores be obtained for each test: the number right and the number attempted but wrong. Similar, but more sophisticated, test scoring procedures have been suggested in the psychometric literature and involve giving a differential fractional score to each of the wrong answers, reflecting the amount of achievement necessary to choose that particular wrong answer--some answers are more clearly wrong than others to a student with partial knowledge. Such scoring has yet to be applied to real evaluation settings, but it will provide greater sensitivity within the particular testing time limits when it becomes feasible.

The primary problem with use of a dichotomous mastery score for each section of a test is that it still leaves unspecified the procedures for summarizing each individual's performance as a single score. The alternative to a single score for each individual, of using instead a multidimensional set of mastery scores for each individual, would necessarily require multivariate statistical procedures in an evaluation, which somewhat increase the computational costs of data analysis and require substantially greater expertise on the part of evaluation data analysts.

Normalized standard scores and percentile scores are conceptually quite similar: they both are obtained as transformations of raw scores to a symmetric distribution. In the case of normalized standard scores, the results are normally distributed; in the case of percentile scores, they are uniformly distributed (that is, in the norm population, the same number of individuals receive each different percentile score). The valid reason that evaluators prefer normalized standard scores over percentiles relates to the validity of using them in standard statistical data analysis procedures. Analysis of variance and all of its variants depend on normality of scores, and percentile scores deviate from normality sufficiently to distort the conclusions reached from the analyses. Occasionally, the argument is heard that normalized scores are "equal interval" scores, meaning that the difference between a normalized score of 10 and 20 is the "same" as the difference between a score of 20 and 30, and that percentile scores are not "equal interval" scores. The grounds for this argument are extremely tenuous. First, there is one sense in which percentiles are equal interval scores: the differences between the 10th and 20th percentiles and between the 20th and 30th percentiles both represent 10% of the population. Second, tne claim that normalized scores are equal interval scores is based on the theory that the achievement test is measuring some underlying factor in the individual that is normally distributed. This theory is, in fact, plausible because of the central limit theorem, which can be paraphrased as saying that anything (e.g., reading achievement) that is the sum of many independent random component factors will tend to be approximately normally distributed. However, the theory that the underlying factor being measured is normally distributed is only plausible, not proven; therefore, any claim that a gain in a normalized achievement score from 10 to 20 represents an equa' amount of learning as a gain from 20 to 30 should be disregarded.

Finally, growth scale scores are similar to normalized standard scores except that growth scale scores add the additional capability of comparison across different levels of a test. Test publishers produce growth scale scores by giving two adjacent levels of a test to the same or matched sets of students to determine which (normalized) score on one level of the test is equivalent to each (normalized) score on the other level. Using this method, a single scale of achievement can be constructed that ranges from first grade through high school.

Of the several methods of assigning numbers to test performance discussed in this issue, some are clearly preferable to others. First, correction for guessing is essential to remove biases engendered by differential tendencies to guess. No matter how explicit the instructions on guessing are (and they are frequently vague), different kinds of children and children in classrooms with teachers of different personality characteristics are going to exhibit different tendencies to guess.

Second, as long as norm-referenced interpretations are to be made or any comparisons involving forms of analysis of variance are to be performed, the scores should be transformed to normally distributed scores (normalized standard scores, normal curve equivalents, or growth scale scores) before entry into analysis.

Third, careful consideration should be given to the use of multivariate analyses of mastery scores for component skills assessed by tests. Using such analyses, it would be possible to go beyond merely concluding that one group learned more than another to reach conclusions about what types of skills were most effectively learned through different treatments.

Finally, grade equivalent scores should be avoided whenever possible.

Analysis

Introduction

The three issues discussed in this section concern the process of transformation of measurements on Title I projects and participants into information relevant to decision rationales. Frequently this is the weakest link in evaluation and therefore a target for challenging a study's usefulness. Establishing the link depends crucially on the identification of research questions or hypotheses for which (1) there are methods, based on tenable assumptions, for deriving answers to the questions from the data, and (2) policy implications of the answers can be deduced in a clear and logical manner.

From a simple point of view, these issues concern the avoidance of pitfalls that can render well-collected data valueless. From a more sophisticated point of view, they concern pitfalls in the overall design of an evaluation. Proper foresight in study design and data collection is needed to prepare for "airtight" analyses and interpretations. Frequently, the key element can be whether the data collection had included a particular item of data that would verify an assumption needed to validate a chosen analysis, so consideration of data analysis prior to development of questionnaires is essential for valid evaluation.

The three issues discussed in this section concern problems that arise when ideal evaluation designs, including random assignment to treatment and control conditions, are infeasible or are otherwise not implemented. These problems can be dealt with in an *ad hoc* fashion for each evaluation, by careful planning and use of statistical expertise; the purpose of the discussions in this section will be both to point out the problems and to suggest methods appropriate for the ad hoc solutions. It is the opinion of the authors, however, that more wholistic solutions, such as changing the framework of comparisons (as suggested under Issue 1) or finding ways to justify more rigorous information-gathering designs, will ultimately be necessary.

The first issue (Issue 8) concerns the conditions necessary for making inferences from a relative comparison between nonrandom treatment and control groups. Each of the methods proposed is based on some set of assumptions, and the discussion will attempt to estimate the reasonableness of these assumptions and to suggest ways of testing them. The most common analytical method used, analysis of covariance, will be described in some detail.

The second issue (Issue 9) concerns the problems that have arisen in attempts to make inferences about the relations of treatment components and costs to effectiveness. That type of information is the most useful information that can be acquired for the purpose of improving the quality of compensatory education, and yet it has usually been gathered as an adjunct to an evaluation more concerned with some other purpose. As a result, many conclusions concerning the relative effectiveness of different methods that have been made in federal studies of compensatory education are highly questionable. The discussion of this issue will attempt to identify the most crucial threats to validity of such conclusions and to suggest ways of dealing with those threats.

The third issue (Issue 10) concerns methods of aggregation of data. Both the sampling units and measurement units affect the meaningfulness of combining data across projects, and the discussion of this issue will attempt to clarify the alternative acceptable aggregation methods and the reasons others are unacceptable.

Issue 8. What are the conditions for valid comparisons between nonequivalent treatment and comparison groups?

This is an important and controversial issue because there are methods for such analyses at hand that appear at first to be valid but have been shown on closer examination to be responsible for distortions in conclusions. In fact, the difficulty of selecting the appropriate analysis has been suggested as grounds for resolving the issue by avoiding comparisons between nonequivalent treatment and comparison groups. Alternatives to such comparisons were discussed under Issue 1. The perspective for the following discussion concerns what to do when one must make such comparisons. In adapting quantitative analysis methods developed for controlled experiments into the area of quasi-experiments in the field, various assumptions on which the methods were based have been violated, and methodologists have recently focused a great deal of attention on ways to weaken the assumptions and still maintain the validity of the methods.

Nonequivalent treatment and comparison groups are any pair of groups for which it is not true that their members might have been assigned to the other group but for a random (or pseudo-random) event. Any method of assignment, such as matched pairs, that is not functionally random will qualify for having the problems discussed below, but the more different the groups are, the more substantial will be biases be that result from violated assumptions. Basically, the purpose of a comparison group is to provide an estimate of how well the treatment group would have performed if it had not had the special treatment. The purpose of each of the methods discussed here is to transform a nonequivalent control group into a group that, except for the treatment, is identical to the treatment group, so that the comparison is possible. This transformation is not necessary in the case of randomly assigned groups, because any differences between such groups will be random, not biased, and therefore they can be statistically accounted for with a high degree of validity.

There are basically four methods for "equating" nonequivalent groups, although there are a number of variants in methods. The four methods are: (1) matching, long denounced but recently revived by Sherwood et al. (1975); (2) gain score analysis, also frequently derogated but recently revived by Kenny (1975): (3) analysis of covariance (ANOCOVA), a powerful analysis tool in experimental psychology but problem-riddled in educational field research and evaluation; and (4) regression analysis. A fifth "method," ignoring the nonequivalence, might be considered for completeness; however, its merits are so inferior to the methods to be discussed as to rule it out of consideration.

Many of the problems to be discussed are present with all four methods; however, the methods are not equivalent. As background, we shall briefly define and list the assumptions of each method.

Matching is relatively simple to describe. It consists of searching for pairs of subjects (e.g., students, classrooms, or school districts), one in the treatment and one in the comparison group, who are as similar as possible on relevant dimensions, deleting all remaining subjects from the analyses to be done, and then performing analyses (e.g., t-tests) as if the groups were randomized pairs (as if you had selected the pairs prior to the treatment and had randomly assigned which was to receive the treatment). The basic assumption of this method that has been questioned in many ways is that the matching is complete, meaning that there is no systematic difference remaining between treatment and control subjects who are matched that could possibly affect their performance. This assumption is clearly false for educational evaluations when matching is on a single dimension: human behavior, and, in particular, the achievement of cognitive skills, is so multiply determined that no single measure can capture all the systematic variance among people capable of affecting later performance. However, in a chapter in the Handbook of Evaluation Research (Struening and Guttentag, 1975), Sherwood, Morris, and Sherwood have investigated the reasonableness of the complete matching assumption if one matches on a hundred or more variables simultaneously; they found matching to be valid in the case of an evaluation study they carried out. A problem with matching on a large number of dimensions is in finding adequate matches. For example, if matching is on 20 dichotomous variables and 10 variables with 5 gradations of level, the number of cells in the population is $2^{20} \times 5^{10} \doteq$ 10 trillion. Even if some variables are moderately correlated, the likelihood of finding 100 matched pairs in a sample of 10,000 treatment and 10,000 control subjects is small. The solution of broadening the gradations (changing from 5 levels to 2 levels, for example), even if it reduces the number of possibilities to a manageable number, is frequently unacceptable because there can then be systematic variation within levels. Suppose, for example, a low economic status group and an (overlapping) high economic status group were matched on just three levels of economic status. There would be a range of status within each of the three levels, and one would expect that at the lowest of the three levels the subjects originally from the low economic status group would be on the average lower than the "matched" subjects from the high economic status group, and so forth. That is, too coarse a match is really not a match at all.

Although matching by itself does not appear to provide an adequate solution to the problem of comparing nonequivalent groups, it may be useful to do in conjuction with statistical methods described below. The bias in statistical correction procedures is least when the groups are most similar. Whenever matching is undertaken, however, possible distortions in conclusions resulting from matching must be considered explicitly. These distortions generally involve some processes that would act differently to cause a particular score on a matching variable to occur in a treatment group than in a comparison group. See Rubin (1973, 1976a, 1976b) for further recent discussions of matching.

Gain score analysis is similarly easy to describe: the method is to create a derived variable ("gain") by subtracting a pretest score from the posttest score and to perform analyses on this derived variable as if the treatment and control groups were randomized. The basic assumption is that pre-existing differences between treatment and control groups, as evidenced by differences on pretests, will not be correlated with later gains. If that assumption were true, then gain scores would be quite appropriate for comparisons in evaluation, because they focus on the effects of the treatment. The frequently noted fact that gain scores have greater random error components (lower reliability) than either pretest or posttest scores is largely immaterial for moderate- or large-scale evaluations, because increasing sample size reduces the importance of random error components. The basic assumption that gains are independent of pre-existing differences is, however, highly questionable in applications to education. Gains are the result of complex combinations of motivational and cognitive processes, and although achievement evidenced at pretest is also dependent on such processes, subtracting the pretest score will not remove the effects of different motivational and cognitive levels on rate of gain between pretest and posttest. Moreover, gains are subject to the statistical artifact that individuals with high pretest scores will tend to have smaller gains because, for some of them, the high pretest scores were "lucky," and conversely for individuals with low pretest scores; that is, regression to the mean is to be expected.

The third method of interest is ANOCOVA. This method is more complicated to describe, although it is conceptually straightforward. Basically, the method is to focus on posttest scores and to hypothesize that the posttest score is a sum of a number of different effects in addition to treatment effect (usually including the level of achievement indicated by a pretest). All the factors (called covariates) that might have effects are measured; then the amount of effect of these factors (their beta weights or regression weights) is estimated from the data; then all the effects due to nontreatment factors are subtracted from each person's posttest scores; finally, the results are analyzed (residuals) as if they were obtained from randomly assigned treatment and control groups.

The basic assumptions of ANOCOVA are:

1. as with other methods, the assumptions needed for the analysis of data from randomized designs, primarily that observations on different subjects are independent of each other, that there is approximately the same possibility of random error in each individual's score, and that random errors are distributed approximately as the normal bell-shaped curve;

2. that the potency of effects of the covariates on posttest scores is the same in treatment and control groups;

3. that except for factors perfectly measured by the observed covariates, the groups are equivalent, that is, indistinguishable from a randomized pair of treatment and control groups; and

4. as with other methods, that the dependent variable can be assumed to be a linear measure of the underlying factor about which one wishes to draw conclusions (e.g., that a particular gain at the high end of a test score continuum has the same meaning as a gain of the same number of units at the middle and lower extremes of the curve).

The first of these four assumptions, as noted, applies to any of the analytical methods. It is included here, however, because ANOCOVA is the only one of the four methods that includes as an integral part what analysis is to be done after groups are "equated."

Figure 5 is included for those who would like an algebraic description of ANOCOVA. It may be ignored without loss of continuity in reading. Most intermediate-level texts on experimental design (e.g., Winer, 1962) include presentations on ANOCOVA.

The fourth method, residual gain score analysis, is quite similar to analysis of covariance, and at times the two have been confused. Residual gain score analysis consists of (1) calculating estimates of each posttest score based on correlations with pretest scores and other covariate factors, (2) calculating residuals by subtracting the estimates from the actual posttest scores, and (3) performing analyses, such as analysis of variance (ANOVA), using the residuals as the variable of interest. Werts and Linn (1970) have shown that residual gain score analysis is based on a statistical model that is a special case of the model underlying ANOCOVA; that is, it requires stronger assumptions than ANOCOVA. It is a reasonable generalization, therefore, that whenever residual gain scores are reported, statistical significance tests should be based on true ANOCOVA, not on the application of ANOVA to the residuals.

Of the four methods, ANOCOVA appears to be generally the best choice for most situations. Although other methods may be appropriate for situations in which particular assumptions are satisfied, ANOCOVA is more general. Thus, it is with dismay that practical evaluators and educators have heard and read the severe attacks on the method by expert methodologists. These attacks have pointed out ways in which the assumptions might be violated in educational evaluations and how they might distort conclusions.

The first major blow to ANOCOVA came from its use in the Head Start evaluation. Campbell and Erlebacher (1970) pointed out problems, while Cicirelli (1969) and Evans (1970) defended the evaluation. Campbell and Erlebacher's presentation included graphic presentations of the way ANOCOVA, when applied without regard to the assumptions underlying it, can systematically bias evaluations and produce just the sort of negative conclusions that the Head Start evaluation arrived at. The problem they identified is now but one of many for ANOCOVA; it was a particular violation of the third assumption, which Campbell and Erlebacher argued would apply to most evaluations of federal education programs. The problem is that ANOCOVA will not correct for all the possible causes of lower achievement in the disadvantaged group, particularly when the pretest contains a portion of random error. This problem and others are discussed later in this section.

Suppose Y_{ij} is the posttest score for individual j in group i of m groups, and X_{ijk} is the kth of p covariate measures for individual j in group i. Then, we calculate a best estimate of Y_{ij} based on the covariates and treatment as

$$\hat{Y}_{ij1} = \hat{\beta}_1 (X_{ij1} - \bar{X}_{j1}) + \ldots + \hat{\beta}_p (X_{ijp} - \bar{X}_{ip}) + \bar{Y}_i$$

where $\hat{\beta}_k = \sum_{i=1}^{m} n_i \,\mathrm{cov}_i(X_k, Y) \Big/ \sum_{i=1}^{m} n_i \,\mathrm{var}_i\,(X_k)$,

n_i is the number of subjects in group i, $\bar{X}_{ik}$ and $\bar{Y}_i$ are the averages of X_k and Y in group i,

$$\mathrm{cov}_i\,(X_k, Y) = \frac{\sum_{j=1}^{n_i} X_{ijk}\, Y_{ij}}{n_i} - \bar{X}_{ik}\, \bar{Y}_i, \text{ and}$$

$$\mathrm{var}_i\,(X_k) = \frac{\sum_{j=1}^{n_i} X^2_{ijk}}{n_i} - \bar{X}^2_{ik}.$$

The residuals, $Y_{ij} - \hat{Y}_{ij}$, represent the error of measurement remaining after the effects of the covariates and the treatment have been accounted for.

The second step is to calculate the best estimate of Y_{ij} based on the same covariates but ignoring group distinctions:

$$\hat{Y}'_{ij} = \hat{\beta}'_i\,(X_{ij1} - \bar{X}_1) + \ldots + \hat{\beta}'_p\,(X_{ijp} - \bar{X}_p) + \bar{Y},$$

where now $\hat{\beta}'_k = \dfrac{\left(\sum_{ij} X_{ijk}\, Y_{ij}\right) \Big/ \sum_i n_i - \bar{X}_k \bar{Y}}{\left(\sum_{ij} X^2_{ijk}\right) \Big/ \sum_i n_i - \bar{X}^2_k}$

and the averages, $\bar{X}_k$ and $\bar{Y}$, are for the total set of subjects.

continued

Figure 5. Algebraic description of ANOCOVA

The residuals, $\hat{Y}_{ij} - Y_{ij}$, represent the error of measurement remaining after the effects of the covariates have been accounted for. If the treatment has no effect, then they should be approximately the same size as the previous residuals calculated. If the treatment is effective, these residuals should be much larger than those previously calculated.

The ANOVA test statistic is

$$\frac{\sum_{ij}\left(Y_{ij} - \hat{Y}'_{ij}\right)^2 \Big/ \left(\sum_i n_i - 2\right)}{\sum_{ij}\left(Y_{ij} - \hat{Y}_{ij}\right)^2 \Big/ \left(\sum_i (n_i - 1) - 1\right)}$$

which is compared to tables of the F-distribution, with

$$\left(m - 1, \ \sum_{i=1}^{m} (n_i - 1) - 1\right) \text{ degrees of freedom.}$$

If the obtained statistic is larger than the table entry for, say the .05 level of significance, the conclusion is that there is at least a 95% probability that the groups differ because of the treatment.

Figure 5. (Algebraic description of ANOCOVA), continued

In a more recent evaluation, the Compensatory Reading Study (Trismen et al., 1975), ANOCOVA was used where the covariates for predicting the posttest score were (1) the pretest score and (2) the square of the pretest score. This means that the estimates can be curved (quadratic) functions of the pretest score--not any possible curve but only simple concave or convex curves. One reason for using the quadratic term is that the levels of pretest scores of compensatory participants and others are different, and curvilinear regression allows for the legitimate possibility of a different regression slope (Assumption 2) between the two groups. See Figure 6 for a pictorial example of such a case. The Compensatory Reading Study's analyses of covariance were plagued with having to reject the analyses because of violations of Assumption 2 (equal regression slopes within different groups). Even with the quadratic term, 44 of 160 critical tests of hypotheses in that study were uninterpretable because of lack of homogeneity of regression slopes between the groups being compared. Lack of homogeneity of regression slopes means that pretest and posttest are more highly correlated in one group (in Figure 6, the compensatory group) than in the other.

There are numerous explanations of differential slopes. Among them are floor and ceiling effects, to be discussed below. The Compensatory Reading Study made great efforts to avoid floor effects, but scatterplots indicated some ceiling effects. Guessing can cause slopes of regressions to vary across the range of pretest scores (i.e., will cause nonlinear regressions); deviations of score distributions from normality will produce nonlinear regressions; and differential growth rates can produce nonlinear regressions. A significant problem with the use of the quadratic term in the ANOCOVA by the Compensatory Reading Study was lack of investigation of the causes of the nonlinearity. A more careful analysis would be likely to suggest a particular type of curve, rather than an arbitrary parabola, and it might even suggest a transformation of the scores that would lead to linear, homogeneous regressions (the Compensatory Reading Study analyzed raw scores, not normalized scores).

The technical summary of the Compensatory Reading Study (USOE, 1976) includes several alternative analyses that produced varying results when applied to the same data. Among them were gain score comparisons and comparisons of relations to a national norm population. Although the results of a the residual gain score analysis carried out by Educational Testing Service

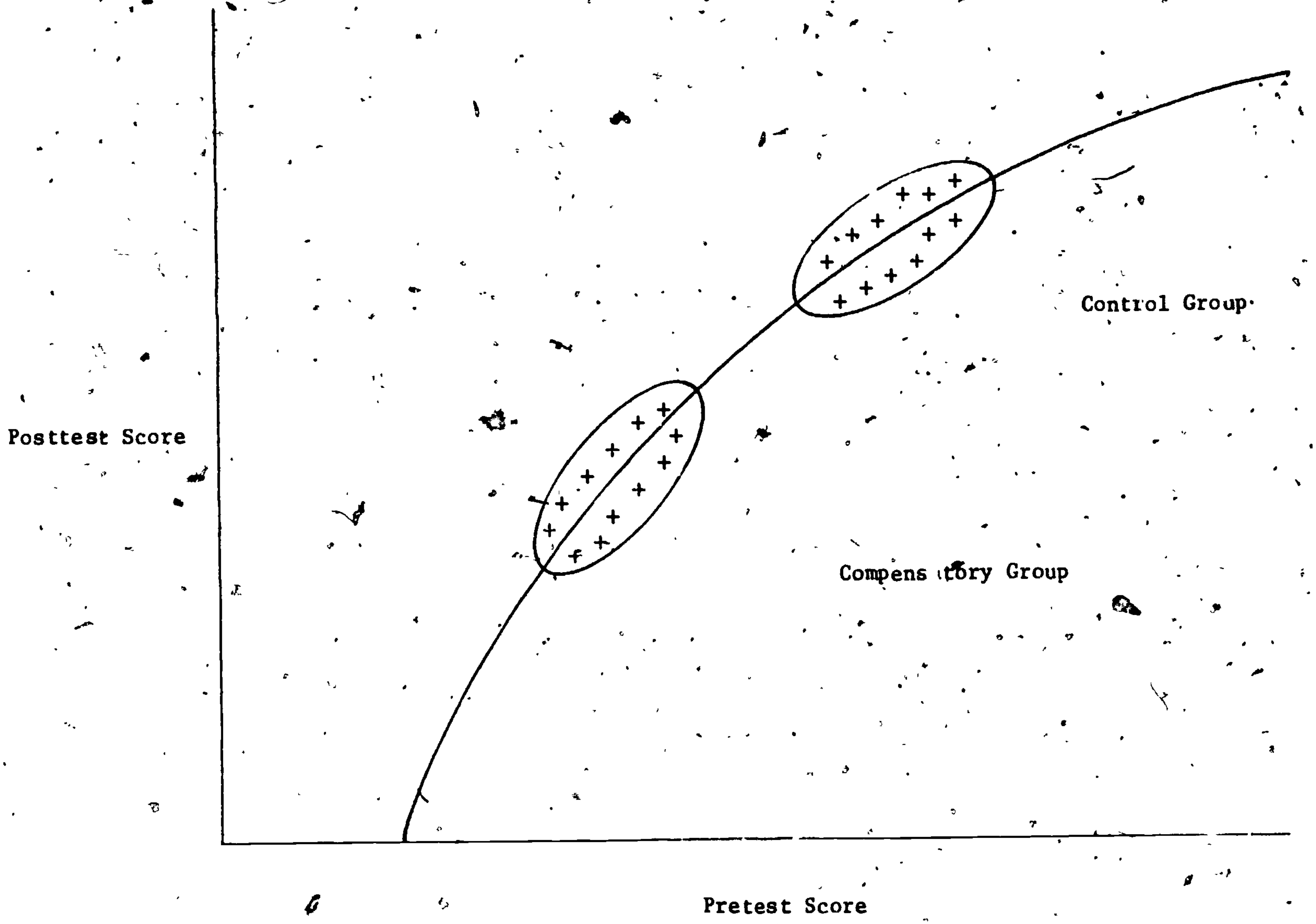


Figure 6. Example of a quadratic regression that fits different slopes for two groups

in that study (referred to in Trismen et al., 1975, as analysis of covariance) slightly favored the noncompensatory groups, the results of the other analyses, carried out by USOE, slightly favored the compensatory reading groups. That different results arose from these different analyses is not helpful for the utility of the study. Ideally, the results should converge to the same conclusion, so the audience could feel confident that the conclusion was independent of the analytical method.

A third use of ANOCOVA in compensatory education evaluation is imminent. The U.S. Office of Education has undertaken to provide technical assistance to state and local education agencies in their efforts to carry out evaluations. As a vehicle for this technical assistance, RMC has developed several evaluation models (Horst, Tallmadge, and Wood, 1975), some of which involve ANOCOVA. "Model C" in that framework involves the use of ANOCOVA for a particular type of nonequivalent treatment and control group. The essential concept of that model is shown in Figure 7. The procedure is to give a pretest and to select for compensatory treatment only those students who fall below some criterion level. Then, after treatment and posttests are complete, the procedure is (1) to calculate the relationship between pretest and posttest based on the control group, (2) to extrapolate thi relationship to predict the treatment group's posttest scores, anu (3) to test whether the treatment group's scores are significantly different from (hopefully above) their predicted levels.

This model, discussed in abstract terms by Kenny (1975) and in more detail by Rubin (1977), cleverly avoids criticisms leveled at other ANOCOVA models in that it does not allow groups to differ in any systematic way not perfectly measured by the pretest. This is accomplished by allowing the teacher no freedom to introduce any other factor besides the pretest score into the determination of who is in the treatment and control groups. Of course, that means that if a teacher used his/her judgment during assignment of students to the compensatory education class, "knowing" that a student could perform better than his/her score indicated or that a student happened to make lucky guesses on the pretest, the results using Model C would be distorted. The cleverness of the model may also be a weakness in another sense: more than any other variant of ANOCOVA, it depends on the assumption that the two-dimensional scatter of pretest and posttest scores conforms to a (bivariate) normal distribution. Although it is straightforward to transform pretest scores and posttest scores separately to a normal distribution (see Issue 7), that does not

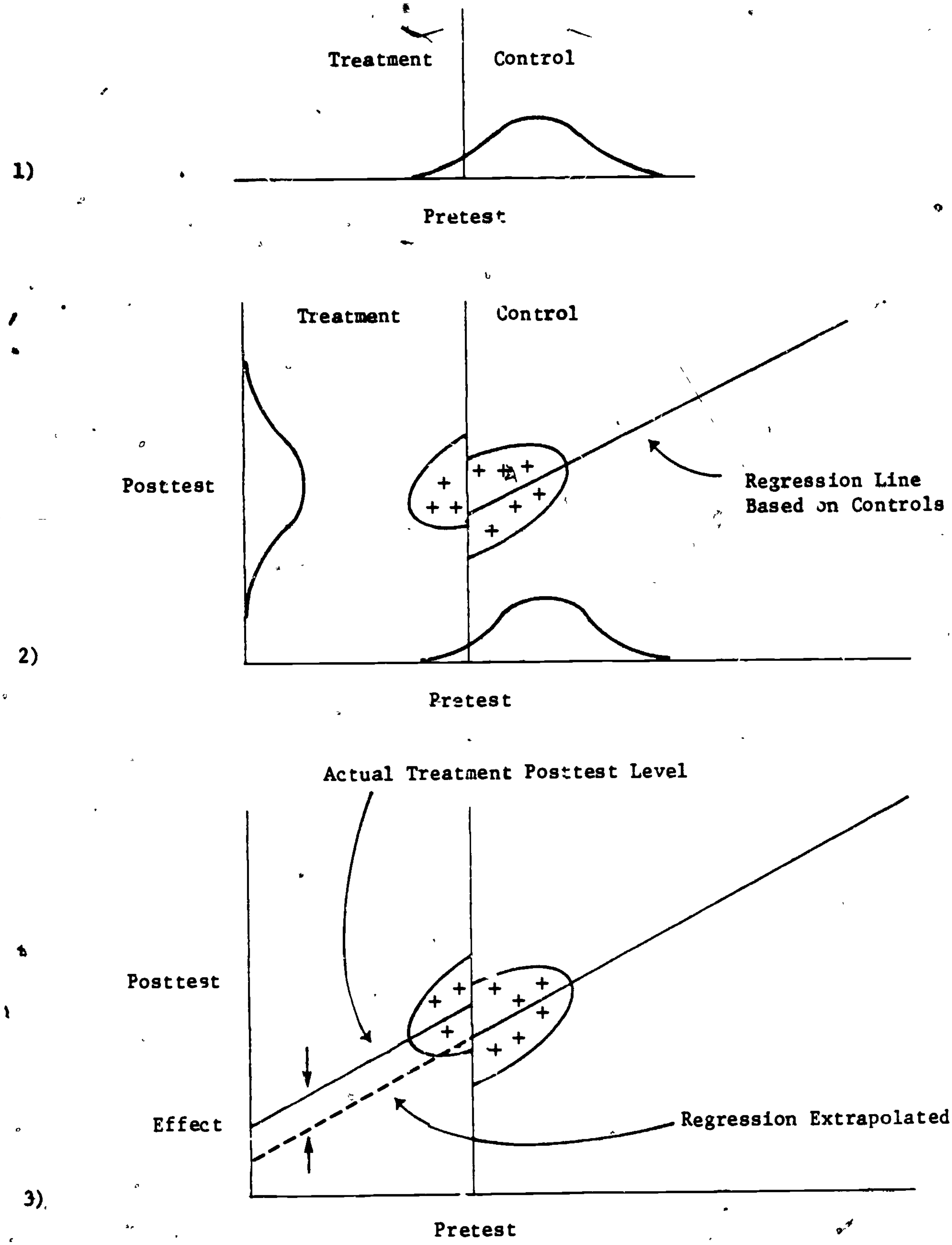


Figure 7. The RMC Model "C" for nonequivalent treatment and control group comparison

ensure that the two-dimensional scatter will be a bivariate normal distribution or that the regression will be linear.

Another problem with this solution is that it fails to address the question of whether the two groups (compensatory and regular instruction) are really fro the same population: it assumes they are, but Campbell and Erlebache- (1970) have argued that they may be different. If you select only according to a pretest, it still may be that you are separating populations that have different achievement expectations. Because the solution appears to be gaining a significant degree of popularity, we digress to describe an example in which selection on the basis of a "pretest" would obviously separate according to populations and would therefore lead to distorted conclusions. Suppose that there were a classroom with 10 English-speaking (Anglo) fifth-graders and 5 non-English-speaking Mexican-American fifth-graders and a third of the class were assigned to a remedial reading program on the basis of an English vocabulary test. With high probability, the Mexican-American children would be given the treatment, and no amount of statistical equating would remove the population effects on a reading posttest; it is just not meaningful o extrapolate from the results of a comparison group to the expected results for a different population. The point is that selecting purely or the basis of a pretest does not ensure that the treatment and comparison groups are alike except for pretest scores.

In order to understand broadly the controversy over ANOCOVA, we need to examine some types of effects that lead to violation of the assumptions of the method. Campbell and Boruch (1975) have discussed six such problems that are well known at present. More problems and variants of the problems and new problems with new variants of ANOCOVA are to be expected. The six problems discussed by Campbell and Boruch are:

1. underadjustment of pre-existing differences;
2. differential growth rates;
3. increases in reliability with age;
4. lower reliability in the more disadvantaged group;
5. test floor and ceiling effects; and
6. grouping feedback effects.

Each of these problems will be dealt with here briefly.

Underadjustment of pre-existing differences violates the third ANOCOVA assumption in that differences remain after the effects of the covariates are partialed out. These underadjustments arise from any systematic rules that lead to assignment to groups other than by a single perfectly reliable measure. The underadjustment arises from the "regression-to-the-mean" artifact in estimating posttest scores in ANOCOVA. Whenever regression is used to estimate scores and the covariate has a random error component, the observed regression line will be less steep than the slope of the underlying relationship (see Figure 8). For example, suppose $X_i = T_i + E_{1i}$ and $Y_i = T_i + E_{2i}$, where E_{1i} and E_{2i} are random error components. Since, except for random error, X and Y are both equal to T, the "true" relationship would logically be $\hat{Y} = X$. However, if the variance of the errors is, say, 10% of the variance of T, then the observed relation will be $\hat{Y} = .909X$. That is not an error of the regression method but rather a theoretical limitation of measurement.

If there are some population differences between those students selected for treatment and controls, such as teachers' judgments of aptitude, that are measured by the pretest but with some small random error, and if that difference has any effect at all on posttest scores that is not reflected in the pretest, the ANOCOVA test statistic will tend to indicate the posttests of the two groups are farther apart than they really are, because ANOCOVA assumes that except for the pretest the groups are completely equivalent. A solution to this problem has been proposed by Lord (1960), Porter (1967), and Porter and Chibucos (1974) and discussed and extended by Campbell and Boruch. The solution involves measuring the reliability of measures used as covariates and then increasing the regression coefficients to correct for the error in the covariate. In our example above, knowing that the variance of errors is 10% of the variance of T, or that the reliability of X is

$$\frac{\text{variance of T}}{\text{variance of T} + \text{variance of E}} = .909,$$

we would divide our observed regression coefficient by the reliability to obtain a hypothesized relation of $\hat{Y} = 1.00X$, which is the true relation. This correction, referred to as "true score analysis," was investigated by Marston and Borich (1977), who found that it tended in some cases to produce too many statistically significant results. St. Pierre and Ladner (1977) investigated the effect of this correction on the results of the Follow-Through evaluation

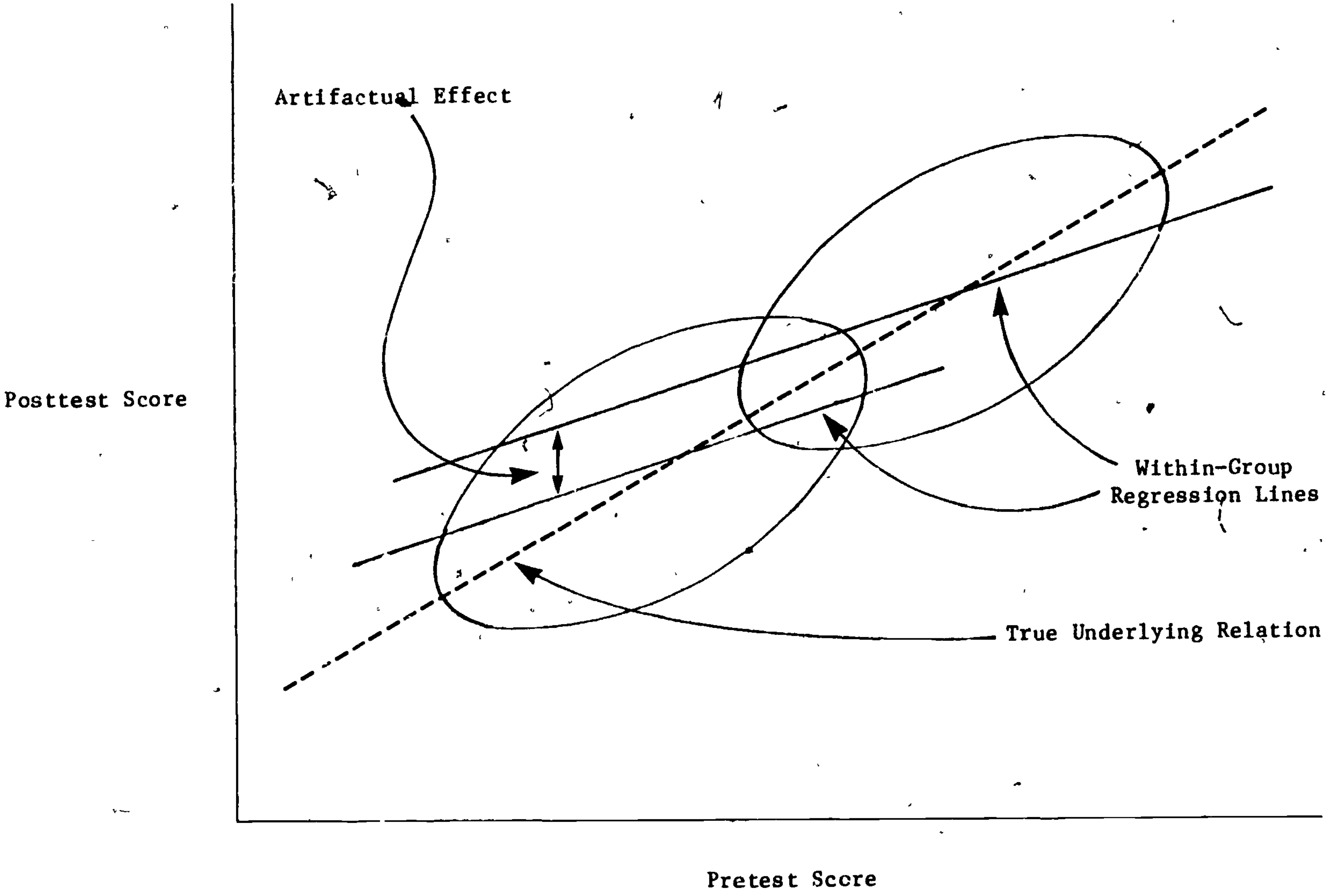


Figure 8. How regression tc the mean affects ANOCOVA

and found that the results did in fact change when the correction was made, so one cannot rely on the easy reply that "it doesn't make much difference anyway."

Differential growth rates are well known to occur in education. One need only look at test publishers' growth scale curves to see that (1) younger children learn faster (e.g., the overlap in scores between first and second graders is less than the overlap between fifth and sixth graders) and (2) children at the lowest percentile levels learn slower than other children. Thus, equating groups on a pretest, whether it is done by matching, by gain score analysis, or by ANOCOVA, will not necessarily equate them on expected growth rate, so the treatment with the fastest learners will be the one that appears most successful. Kenny (1975) has proposed that if one can collect data on expected differential growth rates, use of those data in a standardized gain score analysis would be appropriate.

Increase in reliability with age, which results from the attributes of standardized tests that they tap more true score variance and less random error among older students, has the effect of making scores that are equally far apart on pretest and posttest appear to be more reliably (statistically significantly) different at the time of posttest. Campbell and Boruch point out the need for a model of reliability change so that analyses will be able to correct for this artifact, and they propose such a model, but they note that their "model is still very primitive and oversimplified."

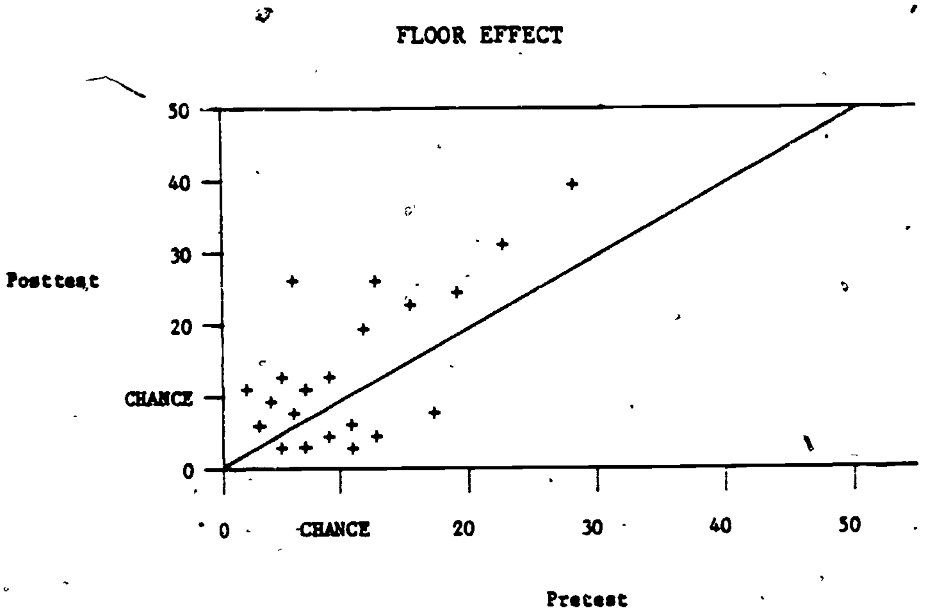
Lower reliability in the disadvantaged group is another way in which Campbell and Boruch suggest that equal true score gains can result in greater observed score gains for one group than for another. The gain, although equal for the two groups, will be less statistically significant for the disadvantaged group.

Floor and ceiling effects can be quite serious, because it is nearly impossible to correct for them after they occur. If a large percentage of students achieved a perfect score on a posttest, it is certain that their gains would be underestimated, but by how much is unknown. Furthermore, for ANOCOVA, the slope of the regression curve of posttest as a function of pretest among the students at the ceiling will be nearly horizontal, because no differences on posttest will be observed for these students although there may be differences at pretest. Therefore, extrapolating linearly to the students of lower ability would put the lower ability students at a disadvantage.

In the case of floor effects, the result of testing will be that gains are underestimated for individuals with pretest levels of achievement much lower than the level that is needed to barely exceed chance performance. Some students will even exhibit "negative learning" because of "lucky" guesses on the pretest. Thus, treatments that are applied to individuals at ability levels lower than those for which the achievement pretest is designed will be much less likely to show systematic gains from pretest to posttest than treatments applied to students in the midrange for the test (see Figure 9).

How might one detect, and correct for, floor effects? Detection is fairly simple. If there are any scores below the chance level, then some floor effects are probably present. Some students may not guess, however, so their scores even though below chance level would not be at the test floor; thus, control of guessing (e.g., encouraging it) is important and, more important, scores should be corrected for guessing, taking into account the number of items attempted, in order to identify floor effects. Correction for floor effects is more difficult, so difficult that the use of "out-of-level" tests specifically to avoid floor effects, such as used in the Compensatory Reading Study (Trismen et al., 1975) is recommended. The problem with choosing a lower test level to fit the achievement range of compensatory education participants' is that regular students are likely to score at the ceiling of that test and comparison using two different tests would rely too heavily on the test publisher's articulation between the levels.

The issue of ceiling effects is somewhat different from floor effects for two reasons. First, the ceiling effects occur in the comparison group, not the treatment group, in compensatory reading programs; and second, ceiling effects are more clearly observable, since the scores are not contaminated by guessing behavior. The first difference is important because the comparison group is taken as the standard against which to compare the treatment, and that means that model parameters, estimated for the comparison group (as in RMC's Model C), will be greatly affected by the ceiling effect. These parameters are the average amount of growth in achievement, the variance of growth scores, and the correlations between pretest and posttest scores. The ceiling effects will lead to underestimation for the comparison group of average gains, variance of posttest scores, and correlations between pretest and posttest scores. These results of ceiling effects will cause linear extrapolation of the relation between pretest and posttest scores from the comparison group to the range



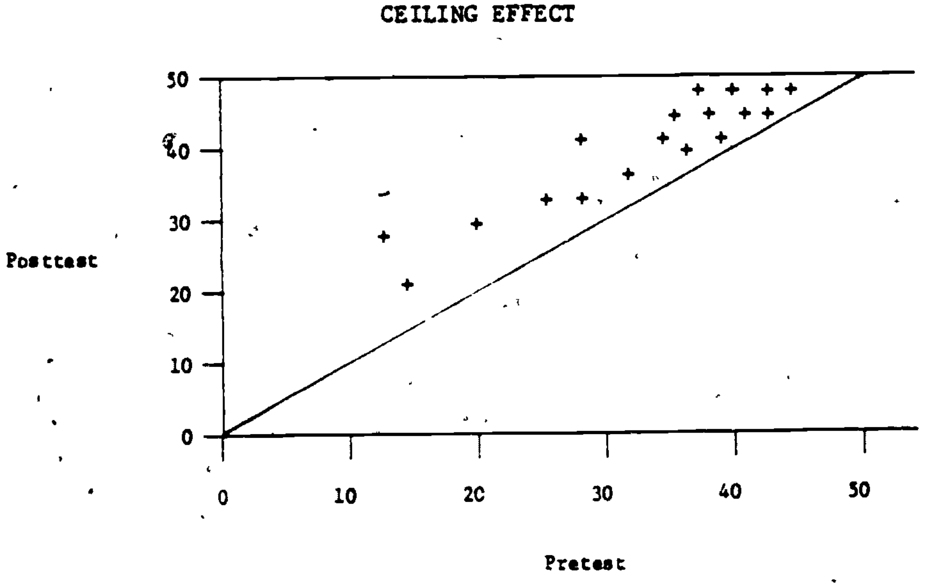


Figure 9. Examples of typical floor and ceiling effects in bivariate distributions

of pretest scores obtained by the treatment group to produce larger expected gains (i.e., a more difficult criterion) than if the ceiling effect were not operating. To deal with this potential problem, the Compensatory Reading Study used a quadratic extrapolation, which has not been well investigated, but is likely to correct (or partially correct or overcorrect) for the ceiling effect.

The detection of ceiling effects is easy: are there any perfect scores? It would be reasonable to correct for ceiling effects by transforming perfect scores upward in order to produce a symmetric distribution, or alternatively, to delete from the comparison group used in the study all students achieving a pretest score higher than the lowest pretest score of a student achieving a perfect score on the posttest. This latter procedure could be slightly refined to account for the possibility of achieving a perfect score by guessing at one or more items. In general, such corrections are more reasonable for ceiling effects than floor effects, because the role of guessing is so much less at the top ~~of~~ the test scale; the higher a student's score, the less will guessing be a contributing factor to that score.

The problems of ceiling and floor effects we have considered pertain particularly to the case of treatment and comparison groups with unequal ability levels. When both groups suffer from identical floor (or ceiling) effects, the problems dissolve into the simple problem of overall lack of sensitivity, which can be avoided by choosing a different test or test level.

Finally, there is a substantive problem of grouping feedback effects. This is the set of effects due to different sets of peer interaction. When compensatory education participants are in a separate environment, they provide an environment for each other that is different from the environment in the regular classroom. This effect cannot be "partialed out" to observe the true instructional treatment, because in a real sense the selection process is part of the total treatment.

In summary, the purpose of methods for comparing nonequivalent treatment and comparison groups is to make them as similar as possible so that differences in outcome can be attributed to the treatment. The weakness of the methods that is most likely to destroy the credibility of conclusions derived from such comparisons is the finding of important pretreatment differences between the groups (or even the argument that there may have been such differences) that were not taken into account in the analyses. Therefore, two important recommendations can be made.

First, the groups should be selected in order to be as similar as possible, maximizing the overlap of similar members. In cases where this is prohibited, as in RMC's Model C, the assumption of the analyses that the treatment and control groups learn according to the same patterns and principles is highly questionable--unless, as Rubin (1977) points out, the evaluator is reasonably certain on the basis of prior knowledge that those patterns are the same. In attempting to match groups, some caution is necessary, however. If matching is achieved partially because of unreliable chance variation (e.g., when matching on a pretest of less than, say, 95% reliability) so that the match would not persist throughout the evaluation, then differential regression to the mean will confound the analyses. Therefore, matching should be made on the basis of reliable measures.

Second, various sources of difference between treatment and comparison groups should be explicitly noted in planning and reporting the study, and measurement of all potential differences and use of those measurements in analyses should be undertaken.

Given that these recommendations are followed, then the use of analysis of covariance, followed by subsidiary analyses to evaluate the distortion in results due to the nonequivalence of the groups, seems appropriate, if randomized assignment is ruled out. Because of the controversy concerning the correction for unreliability of the covariates, that procedure appears questionable at present: it should be used only, as by St. Pierre and Ladner (1977), in conjunction with uncorrected analyses to determine the possible effects of unreliability of the covariate on the results. Improving the reliability of the covariates is preferable; one possibility in the educational evaluation area might be to use the gain (posttest minus pretest) as the dependent variable and the sum of the posttest and pretest scores as a more reliable covariate. This would sacrifice reliability in the dependent variable, which implies merely a loss of precision in results, in order to gain reliability in the covarate, which reduces the bias in the results. The greater reliability of the covariate derives from its being the sum of two measurements of the same construct, and more information corresponds to greater reliability. One might worry that this will confound the analyses because the prctest and posttest are both used in calculating the covariate; however, if gain is the true variable of interest, then that does not matter: knowing the sum of the pretest and posttest scores tells one absolutely nothing about the amount of gain between them (unless floor or ceiling effects are noticeable).

The subsidiary analyses one should plan to carry out when using ANOCOVA on nonequivalent groups include at least: (1) estimation of the reliability of the covariates; (2) demonstration that, on one or more measures not expected to be directly affected by the treatment, partialing out the effects of the covariates does in fact eliminate group differences; (3) testing the functional form of the regression equation by fits to scatter diagrams, both visually and statistically; and (4) whenever alternative explanations of results appear plausible, performing the analyses in different ways in order to demonstrate the range of possible conclusions one could reach based on the data. These types of analyses have not customarily been carried out, primarily because they were not planned for; when they have been carried out, they have added substantially to the credibility of evaluation results. Therefore, it seems important to include plans for such analyses in future evaluation studies.

Issue 9. Under what conditions can one infer relationships of Title I costs and treatments to effectiveness?

Information on the selection of services that maximize the benefits to be derived from various levels of Title I expenditure is, in the long run, the most important information that evaluations can provide. In order to gather that information with adquate validity to provide the basis for widespread selection of treatments, carefully controlled comparisons involving true experimental designs are called for. Correlational data gathered from ongoing projects are subject to great distortion, but these are the data most readily available. The discussion of this issue will point out four kinds of difficulty in making inferences about treatment-effectiveness and cost-effectiveness relationships and will suggest ways of dealing with the difficulties.

The four types of difficulty are (1) in identifving the contributions of Title I, (2) in comparing treatments with different objectives, (3) in identifying what relationship one should study, and (4) in making causal inferences from correlational data. Each of these difficulties has played a role in the design and outcome of Title I evaluations.

The first difficulty, identifying Title I contributions, has two sources: the multiplicity of programs designed to meet objectives similar to the objectives of Title I and the unintended side effects of Title I funds. The first problem is due to the plethora of educational programs at the state and federal levels with overlapping goals. While one can usually identify compensatory education services fairly readily from onsite observation, tracking down what components are paid for by Title I can be well-nigh impossible. Moreover, in most if not all cases, Title I pays only a small portion of the total cost of educating any student, so achievement gains can only tenuously be related to Title I services without careful process analysis. The diversity of sources for educational funds is shown in the surveys by the National Center for Educational Statistics (NCES, 1976, 1976). During the 1971-72 school year, at least eight different federal programs provided funds for reading instruction, with 92% coming from Title I, and during the 1972-73 school year, there were at least ten programs, with 85% coming from Title I. Thus, even though reading instruction is the subject matter most closely related to Title I among federal education programs, other federal programs as well as state and local programs supported significant reading instruction. A report on compensatory education in California in the 1974-75 school year (California State Department of Education, 1976) covered three state programs as well as Title I and found that there were more individual compensatory reading programs at each grade level with Title I plus other sources of funds than with Title I funds alone. Although California is hardly typical, a quote from the summary of that report will give an idea of the complexity of divisions of funds from various sources into various services:

> In ECE [State Early Childhood Education program], 55% of the funds went to pay classified salaries, and 21%. . . for certificated salaries. In ESEA Title I programs, 43% of funds were used for classified salaries and 33% for certificated salaries. In EDY programs [Education for Disadvantaged Youth], 10% of the funds went to pay classified salaries, while 71%. . . for certificated salaries (page 60).

Did EDY programs pay for teachers, and other programs for support personnel? Is there an accounting procedure that makes it simpler for local districts to assign some funds to some services and other funds to other services?

Because of the myriad sources of funds for most of the school districts that receive Title I funds, it is in fact infeasible to obtain estimates of the Title I effects at any reasonable cost--that is, if what is required is an estimate across the nation. The mere fact of the continued existence of Title I and its ramifications in terms of effects on the development of state compensatory education programs and other compensatory education programs makes it impossible at this point, even in theory, to estimate the total Title I effect in most school districts. On the other hand, it may be possible by an intense, in-depth analysis of the budgets and services and impact of Title I within a small number of school districts to estimate what actually was the direct Title I effect. Where Title I contributions are inextricably mixed as the funds from other sources, proportional allocation of the "credit" for benefits would be possible. This is one area in which care must be taken not to allow the need for information to interfere with optimal use of Title I funds, except possibly for a negligible distortion in a few districts randomly selected for special study.

Of more interest than isolation of Title I contributions may be examination of the effects of expenditure variations on whether compensatory education programs of any type work. In fact, for the fundamental purpose of program evaluation, planning for the future, it is not as important to find out what the Title I contribution has been as to find out how to direct Title I expenditures to increase the effectiveness of other projects in the future, that is, to perform a cost-effectiveness analysis.

Many side effects of Title I funding can be imagined, such as increasing the number of jobs for reading aides in impoverished communities. For the purposes of evaluation in terms of children's achievement, however, side effects on children in school are most relevant. The most salient side effect is likely to be enhancement of the scholastic processes for noncompensatory students; providing special resources for educacionally disadvantaged children will in most cases reduce the demands of these children on the regular instructional resources (e.g., teachers' time), allowing greater resources to be devoted to the noncompensatory students. Thus, comparisons between compensatory and regular treatments are less likely to show the benefits of compensatory education than comparisons between matched schools or classrooms in which the "comparison" group has the same membership it would have had if Title I funds were not available (i.e., including educationally disadvantaged children).

Other relevant side effects to be measured in a careful evaluation include (1) filtering of effective compensatory reading methods into the regular curriculum, (2) possible stigma associated with participation in a compensatory treatment, and (3) possible reduction in the effectiveness of regular instruction due to allocation of too much of the available teaching expertise to the teaching of educationally disadvantaged children. The assessment of these and other side effects requires astute onsite observation of the processes occurring during the treatment period. Survey data will almost surely be inadequate.

The second difficulty concerns the multiplicity of objectives of Title I projects. The Elementary and Secondary Act of 1965 was intended to provide services in the schools that would equalize the opportunity of children from low-income areas to e oy a fulfilling education. Many different uses of the money allocated to local districts were attempted. Gradually a few distinctive types of service emerged as most appropriate for Title I expenditures. Table 5 shows a breakdown of expenditures taken from the NCES survey of the 1972-73 school year. Clearly, reading and mathematics have become central. One might envision a future in which the Title I program is divided into subprograms of math instruction and reading instruction; however, there are advantages to comparing the different services within a single framework as well as advantages to analyzing them separately.

One reason for making comparisons across different services is to determine which types of service have broader impact. A service which would result in a child's improvement in several scholastic areas would have apparently greater utility than a service that merely improved performance in a single area. One might guess, for example, that compensatory reading instruction would have broader impact than compensatory social studies instruction; and if they have impact at all, food, health, and counseling services may have the broadest impact. To compare different services, it would seem necessary to determine a vector of criteria for achievement and other potential outcomes and to measure gains from a particular type of service on all these criteria. Thus, one could operationalize the guess that reading is broader than social studies by predicting larger combined total gains in reading, social studies, and mathematics as a result of reading instruction than as a result of social studies instruction.

There are other reasons for comparing different services in the same framework: studies of principles of successful programs in one service area may yield insights into successful methods for other services; critical prerequisites, such as grade level, maturity, and other basic skill achievement, may determine when a particular compensatory instruction is best conducted; and there may be mutually facilitory or inhibitory effects of simultaneous reception of two or more different Title I treatments. Clearly, ananlysis of services aimed at different objectives is worthy of study.

On the other hand, it is quite reasonable for a national evaluation with limited resources to focus on a single type of service, as the Compensatory Reading Study did, rather than to compare mixtures of different services. Data on 200 compensatory reading classes is much more likely to yield results

Table 5

Percentage Expenditures for the Title I Low-Income Area Support Program During the 1972-73 School Year

Direct Services		67%
Reading (English)	38%	
Other English Language Arts	6%	
Mathematics (and Natural Science)	11%	
Other Basic Skills	11%	
Other	1%	
Support Services		31%
Pupil Services	10%	
Fixed Charges	8%	
Other	13%	
Other		2%

Source: NCES (1976)

of defensible validity than data on 100 compensatory reading classes, 50 compensatory mathematics classes, and 50 other compensatory treatments.

The decision of whether treatments with qualitatively different objectives should be included in the same study depends on the particular information needs being satisfied. If a general description of the program is needed, then it seems appropriate to include all treatments, but if information on the effective methods for compensatory education is sought, comparisons should be made only between treatments that have common objectives.

The third difficulty is in the identiiication of the relationship to be studied. Although this may seem obvious, it is not. The important information may not be merely that when variable A is increased, so will variable B be increased. As a practical example, a controversy around 1972 concerned whether there was a "critical mass" of Title I funds that needed to be spent on each participant (e.g., $100 per year or $300 per year) in order to have an impact on his/her achievement. The implications of this issue for policies of concentrating funds on a few of the most disadvantaged children are clear. Although in order to address this question properly, a great deal of seconcary resource availability information is required, it can be approximated by examination of the relationship of per pupil expenditures and achievement across districts.

In order to answer the "critical mass" question, it is necessary to determine whether there is some value of per pupil expenditure such that expenditures above that level have a far greater effectiveness than expenditure below that level, that is, to determine the point of maximum increase in effectiveness plotted as a function of expenditure as in Figure 10. Tallmadge (1973) merely examined the correlation between expenditures and effectiveness to deal with the critical mass question. Of course, his finding of almost no correlation suggests that other analyses would not turn up a critical mass, so the other analyses may not have been warranted for his data.

Another more general question about relational definition, mentioned in the discussion of Issue 1, concerns whether achievement gains are to be treated as equally important across the scale or whether gains which result in students' surpassing a specified proficiency level are to be treated as

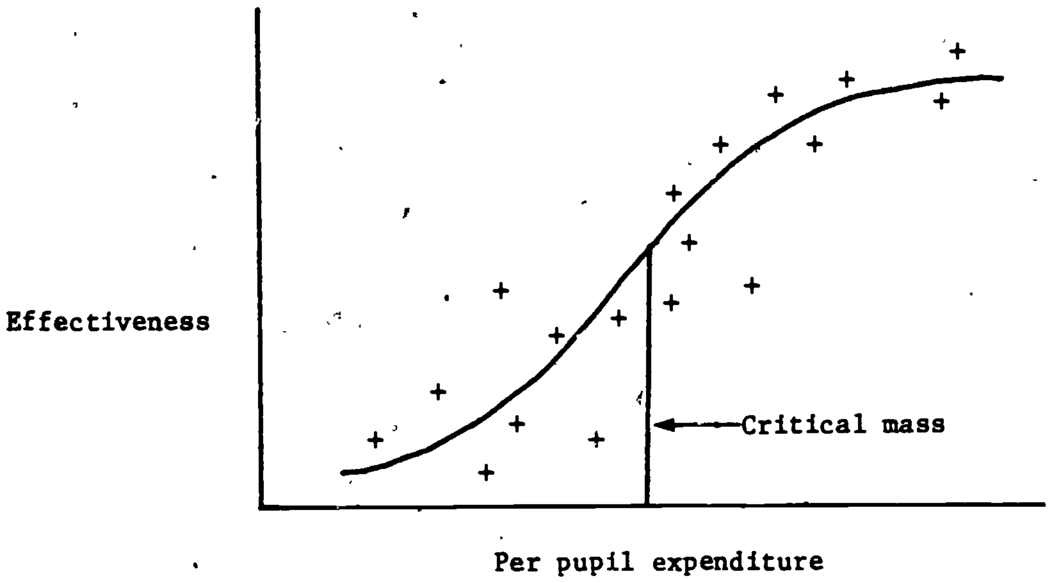


Figure 10. Hypothetical example of a curve-fitting solution for finding a critical mass of expenditure

most important. If a particular method of compensatory instruction focuses on achieving a particular level of achievement for all participants, its production of group average achievement gains is likely to be less than a method that treats all children's gains as equally important, whether they are moderately or severely disadvantaged.

Consider a concrete example. Suppose in a compensatory class there were four students with different learning rates. They required, respectively, 10 hours, 20 hours, 30 hours, and 40 hours to learn a particular amount, say M. Suppose one teacher allots 100 hours as follows: 10 hours to the first student, 20 to the second, 30 to the third, and 40 to the slowest student. Each student would then learn the amount M. Suppose a second teacher allotted 25 hours to each student. The fastest student would learn an amount equal to 2.5 M, the second student 1.25 M, the third student .833 M, and the slowest student just 25/40, or .625 M. The average gain under this teacher would be

$$\frac{(2.5 + 1.25 + .833 + .625)}{4} \; M$$

or about 1.3 M, substantially greater than under the more flexible teacher. The point of this example is that focusing on compensatory class averages instead of, say, class minima, has significant implications for the type of process that will be found to be most effective.

The identification of relations to be assessed in an evaluation depends on (1) clear knowledge about the information needed and the uses to which it is to be put and (2) expertise in translation of verbally stated relations into quantitative calculations.

The fourth difficulty concerns the inference of causal relations from correlational data. If the correlation of a particular instructional process with achievement, across a variety of settings, is positive, then the initial reaction is that the process is effective. There are many other possible explanations of the correlation, however: other events that may have caused both the process to occur and achievement to be high. For example, the process may have been employed in districts containing large numbers of students who would be likely to make higher than average achievement gains, or the occurrence of the process could be merely an indicator of teacher expertise or some other underlying factor that, through other processes, caused achievement to rise.

The solution for making inferences from correlational data is to have a prior, detailed model of the instructional system being observed that includes a chain of related events that lead from processes to effectiveness measures. Each of the events in the chain can then be monitored as well as the occurrence of the process of interest, and finding the predicted chain of correlational results that would explain the correlation of service with effectiveness would rule out most alternative explanations for the correlation. The necessity for a detailed system process model for valid interpretation of correlational data cannot be overemphasized. Without such a model, one should be highly skeptical of all correlational results of compensatory educational evaluations.

In summary, it is our opinion that inferences concerning relations of costs and treatments to effectiveness can be made from surveys and correlational results, but only if a great deal of care and preparation precede such inferences. Inferences from true experimental designs are much more credible, if such designs are feasible. Concerning the other three difficulties discussed above, (1) the isolation of Title I contributions can be very difficult, and for many information needs is not as necessary as identification of compensatory education treatments supported by whatever funding sources; (2) direct comparisons of treatments with qualitatively different objectives is questionable and rarely necessary, although joint study of treatments with different objectives may provide useful results concerning the generality of processes affected by the treatments; and (3) substantially more consideration should be given to the identification of just what relations are to be assessed than has been the case in the past.

Issue 10. How should data be aggregated across projects in Title I evaluations?

The reason for aggregating data across projects is to provide an assessment of the status of Title I throughout the state, region, or the country. This kind of aggregation is clearly necessary for annual reports to Congress and also for general management policy decisions. On the other hand, there are important uses of the local evaluations that do not involve aggregation beyond the district. These are uses, for example, to provide feedback within the district as to what types of services are working and how they are working. Thus, it is quite reasonable for a local district to gather data and analyze, summarize, and report it in a manner that in fact would not allow its being easily aggregated with data from other projects in its state or in the country. In the past, it has been customary to attempt to aggregate all of the local evaluation reports into state evaluation reports, which were then aggregated into a national report to summarize the impact of Title I projects across the country.

There are two aspects of this issue to be dealt with.

1. What are the appropriate units to aggregate across projects?
2. What is the appropriate system for weighting various projects during aggregation?

Major national syntheses of Title I impact (Wargo et al, 1972; Gamel et al, 1975; Thomas & Pelavin, 1976) have been built primarily on annual state reports, and an effect of that has been that conclusions were based on aggregations of grade-equivalent scores, those being the units most frequently reported by the states. This type of national synthesis is a particularly efficient form of national evaluation, because it involves no new collection of data; however, the evaluator has no control over the collection of these data, and as a result both the evaluator and his/her audience have significant doubts as to the data's validity. In the long run, as long as evaluations will be challenged, it is necessary to estimate a minimum level of credibility below which the evaluation is useless and to select an evaluation strategy to ensure that level of credibility. Aggregations of reports generated for some other purpose, while quite useful as corroborative evidence, are dubious as the primary information source. In general, one can say that the collection of data from many projects and their aggregation should follow from an examination of information needs, and then data collection should be carried out in order to satisfy those needs. Of particular importance is the fact that, while every district receiving Title I aid should be carrying out evaluation for its own purposes, the data needed for a national summary evaluation could be supplied by a small random sample of the districts receiving Title I funds as long as that sample is selected in an unbiased and representative manner. Several studies (USOE, 1970; Glass, 1970; NCES, 1975, 1976; USOE, 1976) have based national summaries on a sample of districts.

Let us consider in some detail the measurement units that should be aggregated. Alternative units were discussed under Issue 7. In the past, the rule has most frequently been to transform gains observed or scores observed in particular projects or particular subjects into grade-equivalent gains of month per month and to average these numbers across projects in a state and then across states. Although we might argue about the use of grade-equivalent scores, it is clearly necessary for aggregation that comparable units be entered into the averages for each of the districts that are being aggregated. Certainly raw post-test scores or raw gain scores would not be appropriate for aggregation unless the same test were used throughout the country. But on the other hand, in evaluation studies that do use the same test in all schools, such as in the Compensatory Reading Study, (Trismen et al, 1975), it is more reasonable to average the raw test scores, although normalized standard scores would be preferable. When the scores to be aggregated are from different levels of a particular test, equation for the articulation between the levels must take place (e.g., by use of growth scale scores).

The primary requirements for scores to be aggregatable are (1) that they have the same meaning for all cases that are being aggregated and (2) that the aggregate score have the same meaning for the aggregate group as each score has for the case it represents. Thus, in order to aggregate scores on different tests across projects, it is necessary to aggregate a derived score that expresses the observed performance relative to some expected or national norm performance. Four possibilities are percentile gains, grade-equivalent gains, normalized standard score gains, or percentages of students achieving specified objectives. If any one of these scores is computed for each individual and then aggregated by an appropriately weighted averaging, it will satisfy the second of the two requirements, if it satisfies the first. Percentile, grade-equivalent, and raw gains, however, usually do not have the same meaning for all cases aggregated, if one assumes that normalized standard scores linearly represent the underlying achievement dimension: a particular grade equivalent gain, obtained at the low end of the achievement scale, implies a larger underlying gain than the same grade equivalent gain obtained at higher achievement levels, and a given percentile gain represents a larger "real" gain at the extremes of the scale than in the middle. The validity of the assumption for this argument was questioned and discussed in Issue 7. Also the summary of the Compensatory Reading Study (USOE, 1976) includes an appendix that demonstrates that had that study used grade-equivalent scores, the conclusions would have been seriously distorted. The conclusion arrived at there was that grade-equivalent scores "should never be used in educational evaluations" (page 77, emphasis in original).

Gains in normalized standard scores or normal curve equivalents are especially appropriate for aggregation, because adding them together, unlike other alternatives, does not change their statistical properties: the aggregate score is also normally distributed. Finally, percentages of students achieving specified objectives must be properly weighted to be meaningfully aggregated, and the proper weighting is equivalent to adding numerators and denominators together separately to obtain an overall percentage (e.g., 4 out of 5 in one project [80%] plus 5 out of 10 in another project (50%) yields a total of 9 out of 15 [60%]).

One further note: it is usually not meaningful to transform aggregated units of one type to another type of unit in order to perform further analyses. For example, one might consider transforming the mean grade-equivalent gains reported in annual state Title I evaluation reports into mean normalized standard scores in order to aggregate across states. Theoretically, one could use standard test publishers' tables to make the transformation. However, this transformation would be meaningless, primarily because of the nonlinearity of each derived score as a function of raw scores. The mean of a group of percentile scores is not generally equal to the percentile of the mean of their raw scores, and similarly for grade equivalents and normalized standard scores. Once one has selected a particular measurement unit and performed one level of aggregation, (e.g., calculated a mean), further analysis and aggregation must be in terms of that unit in order to be valid.

Let us consider, now, the problem of weighting the results from various projects in determining an aggregate summary value. Weighting is a method to obtain representative unbiased estimates of population values even though one has a sample with known biases. As mentioned in discussing the various methods in the introduction to the Sampling Section, one can use stratified sampling, sample with different sampling proportions from each of the strata producing a biased sample, and then recombine the data using weights to eliminate the bias. This was done, for example, in the CPIR surveys (NCES, 19765, 1976).

The reasons for sampling in different ratios from various strata are (1) the need for equal precision of estimates in strata of different sizes, (2) differences in the cost of collecting drta from different strata, and (3) effects of sampling units. If one stratum contains 200 schools and another 800 schools, and if one is planning to use a sample of 50 schools both primarily to test for differences between the two strata and secondarily to provide an overall population estimate, then, other things equal, he/she should select 25 schools from each stratum, not the 10 schools in one stratum and 40 schcols in the other stratum needed for representativeness. The population estimate can still be obtained by weighting the schools in the second stratum by four times as much as those in the first stratum (each sampled school in the second stratum represents $\frac{800}{25} = 32$ schools in the population, whereas in the first stratum each sampled school represents $\frac{200}{25} = 8$ schools in the population and $32 = 4 \times 8$).

Different selection ratios based on cost are most noticeable in the follow-up of nonrespondents. Costs may be 10 or even 50 times as great per case in the stratum of nonrespondents as in the stratum or respondents. Thus, the benefit from finding all nonrespondents will rarely justify the costs. Texts on sampling theory (e.g., Raj, 1968) provide formulas for optimal tradeoffs of cost and precision as a function of one's needs for precision.

The third reason for weighting is to reconstruct one population from a sample from another population. For example, if mean achievement levels are available from state reports, they can be used to produce national estimates by weighting each state's achievement level by the number of students in the state.

Briefly, to be explicit, weighting means multiplying each sampled unit's score by the number of units in the population it represents, when calculating means, standard deviations, and so on. In the example of differential sampling from two strata discussed above, if the mean number of students in schools in the first stratum is 150 and for schools in the second stratum it is 300, then the unbiased estimate for the mean for the population of 1,000 schools is not $\frac{150 + 300}{2} = 225$ but rather $\frac{200(150) + 800(300)}{200 + 800} = 270$.

Use of weights, while producing unbiased or nearly unbiased estimates of average values (estimates that tend to be the same as the population value in the long run), also reduces the effective sample size. For the example, the 50 schools produce a weighted estimate of the mean with a standard error equal to an unweighted sample of 37 schools.* Thus, care must be taken not to be too extreme in use of differential weighting in stratified sampling. It should be apparent also that appropriate weighting is impossible if the differential selection ratios are not known.

In summary, the most important problems for aggregation are (1) to ensure that throughout the aggregation process the same measurement units are aggregated and (2) to ensure that the knowledge of different stratum selection ratios is available for use in weighting results appropriately. The measurement unit that is subject to the fewest criticisms appears to be the normalized standard score unit (one example of which is the normal curve equivalent). In any case, in performing an aggregation using any unit and weighting procedure, an analyst needs primarily to address

* If each of n sampled units, u_i, has a weight, W_i, the effective sample size is $\left(\sum_i W_i\right)^2 \Big/ \sum_i \left(W_i^2\right)$. For the case of equal weights throughout, this is equal to n; otherwise it is less than n.

the questions of whether the aggregate score means the same thing for the aggregate group as each individual member's score means for the individual member and whether a particular score means the same thing for each member who might obtain it.

Summary

In this document, we have attempted to answer the question of "What has been learned about evaluation methodology from the decade of compensatory education?" During that decade, tens of millions of dollars have been spent on educational evaluation, and partly because of the political significance of the information produced by the studies, substantial efforts have been undertaken to identify the methodological problems that can undermine the validity of evaluation. From the resulting discussions and controversies, which can be expected to continue, the most positive outcome has been the recognition of the need for further development of evaluative expertise and the expenditure of effort by capable researchers to satisfy that need. The recommendations for evaluation methodology made previously in this document and reiterated in this section are not merely those of the authors, but rather the authors' interpretations of recommendations made by a large number of researchers in this field. Although many of the recommendations remain controversial in 1977, most, we believe, reflect the general consensus among expert evaluators that greater efforts must be made to gather less information more validly.

We deliberately avoided defining "evaluation" explicitly in this document because to do so in any useful way would preclude from consideration studies that are only tangentially evaluative, in this case, of compensatory education. Rather, we focused on the methodology of information gathering, noting that the use of information to test rationales for decisions is common motivation for its being gathered and an important determinant of decisions concerning methods to be used. The issues discussed pertain to four phases of information gathering: design, sampling, measurement, and analysis.

Design

The two design issues discussed did not compare experimental, quasi-experimental, and pre-experimental designs at great length, as was adequately done by Campbell and Stanley (1963). They focused instead on two more global problems: (1) whether quasi-experimental designs could be feasible and what alternatives to quasi-experimental designs might be appropriate for compensatory education evaluation; and (2) whether conditions called for longitudinal data collection paradigms. The major recommendations made concerning design are the following.

Recommendation 1. Future evaluations of the impact of compensatory education should include comparisons of participating children's achievement against a priori, or absolute, standards of expected achievement as wel_ as, or instead of, relative comparisons against the performance of statistically equated comparison groups.

Recommendation 2. When evaluations must provide information based on comparisons between groups, greater effort should be made to find ways of selecting and assigning students to these groups randomly, so that the many problems with statistical equating can be avoided. Several methods for increasing the political feasibility of randomization were discussed. Recommendations for proceeding when a relative comparison against a nonequivalent comparison group is mandatory are discussed in the section on analysis.

Recommendation 3. Individual student achievement gains should be measured for intervals of whole years to avoid distortions that occur from testing twice in the same classroom setting; fall-to-spring gains usually greatly overestimate gains observed over whole year periods.

Recommendation 4. Conclusions based on pretest-posttest gains should not be compared to published norms without taking into account that the children being assessed are taking the test (in parallel forms) twice, whereas the norm group took the test only once, and other test administration artifacts.

Recommendation 5. Teachers' retrospective judgment of children's gains should be disregarded for the purposes of program evaluation; however, teachers' observations recorded during a treatment period can be valuable.

Recommendation 6. Long-term longitudinal studies, making use of overlapping cohorts where possible, are necessary for ultimate impact evaluation of Title I.

Recommendation 7. As a corollary, any evaluations of Title I undertaken without funding for long-term longitudinal data collection should nevertheless take inexpensive steps to ensure that th- data base can later be used as the first stage of a longitudinal study.

These recommendations are made because it is the authors' belief that they would contribute to the improvement of the effectiveness with which education evaluation funds are spent. That they are not completely novel is evidenced by the fact that the design of the current Sustaining Effects Study being carried out by System Development Corporation for the U.S. Office of Education conforms to them more closely than did earlier studies.

Sampling

The two issues dealing with selection of projects or other units for observation that were discussed are substantially less controversial than the other issues in this document, possibly because of the ease of finding compromise solutions (e.g., a medium sized sample) as well as because the theory of sampling is quite extensively developed. The issues discussed relate to the aspects of representativeness and size of samples. The following are the major recommendations that we believe should be made on these topics.

Recommendation 8. The use of quantitatively representative samples should be limited to instances where the information need is for quantitative estimates of program operating characteristics; in other cases, such as testing hypotheses about relationships, other sampling methods are more efficient.

Recommendation 9. The needs for data analysis should be considered in deciding upon the primary sampling units, and great caution should be used in drawing inferences about units other than the primary sampling units. Although valid inferences about student processes can be made when the primary sampling unit is the classroom, it is also very easy to make invalid inferences in that situation.

Recommendation 10. Although there are methods for explicitly deriving needed sample sizes from information precision requirements, the value of precision of information for testing decision rationales is as yet only vaguely understood, so within broad limits the increased costs for large samples may be better spent on more careful study, and therefore more valid information, on smaller samples.

The main theme of these three recommendations is that sampling plans cannot be developed independently from other aspects of information gathering. Greater flexibility in sampling strategies than has been the custom in compensatory education evaluations is called for.

Measurement

The discussion of measurement issues was limited to the measurement of impact on children, primarily on their cognitive achievement. The validity of measurement has undergone the most severe scrutiny of any of the processes in evaluations of compensatory education, possibly because the ways in which measurement can distort reality are more generally understandable than the ways in which sampling or analysis can distort reality, or possibly because of the fact that different ethnic groups obtain different average scores on cognitive achievement tests. The three levels of issue concerning measurement, which provided the structure for that section of the document, are (1) selection of constructs to measure, (2) choice between norm-referenced and criterion-referenced tests, and (3) selection of measurement units in which to record test performance. The major measurement recommendations made are the following.

Recommendation 11. Until more is known about the relations between noncognitive and cognitive gains, measures of noncognitive gains should be used only as supplements to measures of cognitive gains in the evaluation of compensatory education impact.

Recommendation 12. Until more is known about the relations of component skills (e.g., decoding, memory) to overall skills (e.g., reading ability), measures of the component skills should be used only as supplements to measures of overall skills in compensatory education evaluation.

Recommendation 13. Achievement data in compensatory education evaluation should be interpreted in terms of models of cognitive growth processes. In order for this to occur, further research on basic skills is necessary, and the results of that research and existing research must be adapted for use in evaluation studies.

Recommendation 14. Norm-referenced tests should not be used in program evaluation unless the evaluator takes into account the problems in using those tests (eight problems are discussed in this document); in any case, using published norms as the "comparison group" in a relative comparison is highly questionable.

Recommendation 15. Criterion-referenced tests should be seriously considered for use in program evaluation; the most difficult problem to be solved in their use in large scale evaluations is how to aggregate results related to different local treatment objectives.

Recommendation 16. Test publishers should be encouraged in their efforts to provide tests that are both explicitly criterion-referenced and also norm-referenced--these attributes do not conflict.

Recommendation 17. Achievement test scores should always be corrected for guessing when used in program evaluation, based on the number of items each student attempted. This recommendation is made even though it virtually eliminates the possibility of evaluation based on comparing scores on published tests with norms tables.

Recommendation 18. Because of the great heterogeneity of skill levels assessed in compensatory education evaluation, standardized tests sensitive to substantially wider ranges of ability level should be developed; these may require branching processes or differential wrong-response scoring in order to be efficient.

Recommendation 19. Especially when analyses are to be done that assume a normal distribution of scores, but also in other cases, scores should be translated to normalized scores (e.g., normal curve equivalents) as preparation for analysis.

Recommendation 20. Multivariate analysis of vectors of proficiency or mastery scores on sets of component skills should be given serious consideration for program evaluation.

Recommendation 21. Grade-equivalent scores should be avoided.

Analysis

The analytical issues in compensatory education evaluation have drawn the greatest interest of theoretical methodologists. Dealing with these issues provides a useful direction for methodological research, which is also intellectually intriguing. Although three analytical issues were discussed in this document, by far the major interest has been in the first--how to compare the performance of a priori nonequivalent treatment and control groups so that differences can be attributed to the treatment. The other two issues discussed concern the inference of relations (e.g., between treatment processes and effectiveness) from correlational data and the aggregation of data across higher level sampling units. The major recommendations we make on these three issues are the following.

Recommendation 22. Without resorting to unreliable measures, treatment and comparison groups should be selected to be as similar as possible, even when they cannot be randomly assigned.

Recommendation 23. A comprehensive consideration of potential differences between treatment and control groups (prior to treatment) should be a part of evaluation planning and measurements of potential differences between groups on variables relate to performance should be undertaken.

Recommendation 24. Uncorrected, straightforward analysis of covariance is a reasonable method for carrying out comparisons of nonequivalent groups, but only if supplemented by subsidiary analyses that investigate among other things: (1) the reliability of covariates, (2) the residual nonequivalence after partialing out the effects of covariates, (3) the functional form of the regression function, and (4) the change in conclusions that would result if any major untestable assumptions were violated.

Recommendation 25. Whenever causal relational inferences are to be made from quasi-experimental or correlational data, a system model that includes a chain of events that underlies the relation is required, and measurement of at least a subset of the intervening variables is necessary to rule out alternative explanations of the correlation.

Recommendation 26. If scores are to be aggregated across different units (e.g., districts, states, or regions), it is essential that the same measurement unit be used in all cases; if the statistics are in noncomparable units, summaries of summary statistics cannot be made meaningful by statistical manipulation.

Recommendation 27. Information about sampling ratios in different strata must be used in order to obtain unbiased total population estimates using differential stratum weights.

As mentioned before, these recommendations range from obvious to controversial, depending on the reader's viewpoint. Any attempt at synthesis, which this is, cannot explore the details of any particular issue as thoroughly as would an investigator who focused his or her efforts on a single issue; at some point in the not too distant future, many of the issues will be substantially clarified because of the focused efforts of qualified methodologists.

In addition to the limitation in thoroughness, this document is limited in breadth in that not all of the methodological issues potentially relevant to compensatory education evaluation could be discussed. Omissions we feel most unhappy about include a discussion of the Bayesian approach to data analysis, a presentation of quantitative methods for assigning values to program outcomes, an exploration of alternative concepts of basic skills development, a consideration of the external validity of laboratory experiments, and a discussion of issues related to program cost estimation. The issues discussed in this document are, however, the most critical methodological issues for T!tle I evaluation, in our opinion.

In conclusion, the state of the art in educational evaluation has changed dramatically from the situa on ten years ago when the TEMPO study (Mosbaek, 1968) set out to test policy rationales by estimating linear regression coefficients. Much of the effort in that decade has shown the need for further effort to develop evaluation methodology to a level that researchers and policymakers will both find pleasing. New compromises must be found where conflicting values preclude simple solutions (e.g., randomized designs). A primary purpose of this document has been to suggest a few paths to follow in searching for those compromises.

REFERENCFS

Abelson, W. D., Zigler, E., & De Blasi, C. Effects of a four-year Follow-Through program on economically disadvantaged children. Journal of Educational Psychology, 1974, 66, 756-771.

Boruch, R. F. On common contentions about randomized field experiments. In G. V. Glass (Ed.), Evaluation Studies Review Annual, 1976, I.

Briggs, P. G. A perspective on change: The administration of Title I of the Elementary and Secondary Education Act. Report to DHEW/ASPE under contract #HE-OS-72-224. Washington, D.C.: The Planar Corporation, October 1973.

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

Campbell, D. T., & Boruch, R. F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), Evaluation and Experiment. New York: Academic Press, 1975.

Campbell, D. T., & Erlebacher, A. E. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Helmuth (Ed.), Compensatory education: A national debate. The disadvantaged child (Vol. 3), New York: Brunner/Mazel,1970.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Cicirelli, V. G., et al. The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development. Contract #689-4536 between Westinghouse Learning Corporation/Ohio University and Office of Economic Opportunity. Washington, D.C.: Office of Economic Opportunity, 1969.

Cochran, W. G. Sampling techniques (2nd edition). New York: Wiley, 1963.

Cohen, D. D., & Garet, M. S. Reforming educational policy with applied research. Harvard Educational Review, February 1975, 45(1), 17-43.

Coles, G. J., & Chalupsky, A. B. Innovative school environments and student outcomes. Final Report, Project LONGSTEP, Vol. II (AIR-21400-9/76-FR-II). Palo Alto, Calif.: American Institutes for Research, September 1976.

Conner, R. F. Selecting a control group: An analysis of the randomization process in twelve social reform programs. Evaluation Quarterly, May 1977, I(2) 195-244.

Coulson, J. C., Ozenne, D. G., Bradford, C., Doherty, W. J., Duck, G. A., Hemenway, J. A., & Van Gelder, N. C. Emergency School Aid Act national evaluation: The second year of Emergency School Aid Act (ESAA) implementation. (TM-5236/009/0005). Santa Monica, Calif.: System Development Corporation, 1976.

Dienemann, P. F., Flynn, D. L., & Al-Salam, N. An evaluation of the cost-effectiveness of alternative compensatory reading programs. Volume I: Cost analysis. (UR-231) Bethesda, MD: RMC Research Corporation, 1974.

Dyer, H. S., Linn, R. L., & Patton, M. J. A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. American Educational Research Journal, 1969, 6, 591-605.

Eddington, A. The philosophy of physical science. Ann Arbor, Mich.: Ann Arbor Paperbacks, University of Michigan Press, 1958.

Edwards, W., Guttentag, M., & Snapper, K. A decision-theoretic approach to evaluation research. In E. L. Struening & M. Guttentag (Eds.), Handbook of Evaluation Research, Volume I. Beverly Hills, Calif.: Sage Publications, 1975.

Evans, J. W., & Schiller, J. How preoccupation with possible regression artifacts can lead to a faulty strategy for the evaluation of social action programs: A reply to Campbell and Erlebacher. In J. Helmuth (Ed.), Compensatory education education: A national debate. Volume 3, The disadvantaged child. New York: Brunner/Mazel, 1970.

Flanagan, J. C. Units, scores, and norms. In E. F. Lindquist (Ed.), Educational measurement. Washington: American Council on Education, 1951.

Floden, R. E., & Weiner, S. S. Rationality to ritu : The multiple roles of evaluation in governmental processes. Occasional paper of the Stanford Evaluation Consortium. Stanford, Calif.: Stanford University, 1976.

Gamel. N. N., Tallmadge, G. K., Wood, C. T., & Binkley, J. L. State ESEA Title I reports: Review and analysis of past reports, and development of a model reporting system and format. (UR-294) Mountain View, Calif.: RMC Research Corporation, October 1975.

General Accounting Office (GAO). Report to the Congress by the Comptroller General of the United States: Assessment of reading activities funded under the federal program for educationally deprived children. Washington, D.C.: U.S. Department of Health, Education and Welfare, Office of Education, 1975.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement (2nd edition). Washington, D.C.: American Council on Education, 1971.

Glass, G. V. Data analysis of the 1968-69 survey of compensatory education (Title I). Final Report. Boulder, Colorado: University of Colorado, Laboratory of Educational Research, August 1970.

Glass, G. V., Peckham, P. D., & Sanders, J. R. Consequences of failure to meet assumptions underlying the analysis of variance and covariance. Review of Educational Research, 1972, 42(3), 237-288.

Gordon, E. W., & Koutrelakos, J. Utilizing available information from compensatory education and surveys. Final Report. New York: Teaching and Learning Research Corporation, June 1971. (ERIC Document Reproduction Service No. ED 055 664)

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. Sample survey methods and theory. New York: Wiley, 1953.

Horst, D. P. Analysis of school projects for the development of project information packages (PIPs). Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Horst, D. P., Tallmadge, G. K., & Wood, C. T. A practical guide to measuring project impact on student achievement. (017-080-01460) Washington, D.C.: U.S. Government Printing Office, 1975.

Kaskowitz, D. H., & Norwood, C. R. A study of the norm-referenced procedure for evaluating project effectiveness as applied in the evaluation of project information packages. Research memorandum for U.S. Office of Education, Office of Planning, Budgeting, and Evaluation, Contract #OEC-0-74-9256; SRI Project URU-3556. Menlo Park, Calif.: Stanford Research Institute, January 1977.

Kenny, D. A. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. Psychological Bulletin, 1975, 82(3), 345-362.

Kosecoff, J., & Fink, A. The feasibility of using criterion-referenced tests. In Study of the sustaining effects of compensatory education on basic skills: Measures of student growth for the Sustaining Effects Study. TM-5693/003/00. Santa Monica, Calif.: System Development Corporation, 1976.

Lord, F. M. Large-scale covariance analysis when the control variable is fallible. Journal of the American Statistical Association, 1960, 55, 307-321.

Marston, P. T., & Borich, G. D. Analysis of covariance: Is it the appropriate model to study change? Austin, Texas: University of Texas at Austin, Research and Development Center for Teacher Education, 1977.

McLaughlin, D. H. Title I, 1965-1975: A synthesis of the findings of federal studies. Final Report to NIE under contract #400-76-0129. Palo Alto, Calif.: American Institutes for Research, 1977.

McLaughlin, M. Evaluation and reform: The Elementary and Secondary Education Act of 1965, Title I. Cambridge, Mass.: Ballinger, 1975.

McLaughlin, M., et al. The effects of Title I, ESEA: An exploratory study. Cambridge, Mass.: Center for Educational Policy Research, Harvard University, 1971. (ERIC Document Reproduction Service No. ED 073 216)

Mosbaek, E. S., et al. Analysis of compensatory education in five school districts: Summary. Santa Barbara, Calif.: TEMPO, General Electric Company, 1965 (mimeo).

National Center for Education Statistics (NCES). Federally aided programs operated by local education agencies for elementary and secondary schools: National estimates of pupil participation, staff, and expenditures, 1972. (NCES 75-303) Washington, D.C.: U.S. Government Printing Office, May 1975.

National Center for Education Statistics (NCES). Pupil participation, staffing, and expenditures in federally aided programs operated by school districts, 1973. (NCES 76-300) Washington, D.C.: U.S. Government Printing Office, 1976.

National Institute of Education (NIE). Evaluating compensatory education. Interim Report. Washington, D.C.: U.S. Department of Health, Education, and Welfare, National Institute of Education, December 1976.

National Opinion Research Center (NORC). Southern schools: An evaluation of the effects of the Emergency School Assistance Program and of school desegregation, Volume I. (NORC Report No. 124A) Chicago: University of Chicago, National Opinion Research Center, October 1973.

Pelavin, S. H., & David, J. Evaluating long term achievement: An analysis of longitudinal data from compensatory education programs. (EPRC 4537-15) Menlo Park, Calif.: Stanford Research Institute, Educational Policy Research Center, March 1977.

Porter, A. C. The effects of using fallible variables in the analysis of covariance. Unpublished doctoral dissertation, University of Wisconsin, 1967. (University Microfilms No. 67-12, 147)

Porter, A. C., & Chibucos, T. R. Selecting analysis strategies. In G. D. Borich (Ed.), Evaluating educational programs and products. Englewood Cliffs, N.J.: Educational Technology Publications, 1974.

Raj, D. Sampling theory. New York: McGraw-Hill, 1968.

Reichardt, C. S. The statistical analysis of data from the nonequivalent control group design. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

Rubin, D. B. Matching to remove bias in observation studies. Biometrics, 1973, 29, 159-183. (Correction note 30, p. 728)

Rubin, D. B. Multivariate matching methods that are equal percent bias reducing. I: Some examples. Biometrics, 1976, 32, 109-120. (Correction note, p. 955) [a]

Rubin, D. B. Multivariate matching methods that are equal percent bias reducing. II: Maximums on bias reduction for fixed sample sizes. Biometrics, 1976, 32, 121-132. (Correction note p. 955) [b]

Rubin, D. B. Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics, March 1977, 2(1), 1-26.

Ryan, S. (Ed.) A report on longitudinal evaluations of preschool programs. Volume 1: Longitudinal evaluations. Washington, D.C.: Department of Health, Education, and Welfare, Publication No. (OHD) 74-24, 1974.

St. Pierre, R. G., & Ladner, R. Correcting covariates for unreliability: Does it lead to differences in an evaluator's conclusions? Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Scriven, M. The methodology of evaluation. In Tyler, R. W., Gagné, R., & Scriven, M. (Eds.), Perspectives on curriculum evaluation. AERA Monograph Series on Curriculum Evaluation, No. 1. Skokie, Illinois: Rand McNally, 1967.

Scriven, M. Evaluation bias and its control. In G. V. Glass (Ed.), Evaluation Studies Review Annual, 1976, I, 119-139.

Seitz, V. Long-term effects of early intervention: The New Haven Project. Paper presented at the annual meeting of the American Association for the Advancement of Science, Denver, Colorado, February 23, 1977.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. Self-concept: Validation of construct interpretations. Review of Educational Research, 1976, 46(3), 407-441.

Shaycoft, M. F. A guide to the development, evaluation, and use of criterion-referenced tests. (AIR-50700-8/76-FR) Palo Alto, Calif.: American Institutes for Research, 1976.

Sherwood, C. D., Morris, J. N., & Sherwood, S. A multivariate, nonrandomized matching technique for studying the impact of social interventions. In E. L. Struening & M. Guttentag (Eds.), Handbook of evaluation research, Volume 1. Beverly Hills, Calif.: Sage Publications, 1975.

Spady, W. G. Competency-based education: A bandwagon in search of a definition. Educational Researcher, 1977, 6(1), 9-14.

Stake, R. E. The countenance of educational evaluation. Teachers College Record, 1967, 68, 523-540.

Stearns, M. S. Evaluation of the field test of Project Information Packages. Volume I, Summary Report. Menlo Park, Calif.: Stanford Research Institute, 1977.

Struening, E. L., & Guttentag, M. (Eds.). Handbook of evaluation research, Volume 1. Beverly Hills, Calif.: Sage Publications, 1975.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., & Provus, M. M. Educational evaluation and decision making. Bloomington, Indiana: Phi Delta Kappa, 1971.

System Development Corporation. Policy questions addressed by the Sustaining Effects Study. (TM-5693/002/00) Santa Monica, Calif.: Systems Development Corporation, September 1976.

Tallmadge, G. K. An analysis of the relationship between reading and mathematics achievement gains and per-pupil expenditures in California Title I projects, fiscal year 1972. Final Report, Contract No. OEC-0-72-5179. Palo Alto, Calif.: American Institutes for Research, 1973.

Thomas, T. C., & Pelavin, S. H. Patterns in ESEA Title I reading achievement. (4537-12) Menlo Park, Calif.: Stanford Research Institute, March 1976.

Thorndike, R. L. Regression fallacies in the matched groups experiment. Psychometrika, 1942, 7, 85-102.

Trismen, D. A., Waller, M. I., & Wilder, G. A descriptive and analytic study of compensatory reading programs. Final Report, Volume 1. (PR 75-26) Princeton, N.J.: Educational Testing Service, December 1975.

U.S. Office of Education (USOE). Statistical report, fiscal year 1968: A report on the third year of Title I Elementary and Secondary Education Act of 1965. U.S. Office of Education, 1970.

U.S. Office of Education (USOE). A study of compensatory reading programs: Technical summary. Washington, D.C.: U.S. Department of Health, Education, and Welfare, Cffice of Education, Office of Planning, Budgeting, and Evaluation, Elementary and Secondary Programs Division, 1976.

Wargo, M. J., Campeau, P. L., & Tallmadge, G. K. Further examination of exemplary programs for educating disadvantaged children. Final Report. Palo Alto, Calif.: American Institutes for Research, July 1971.

Wargo, M. J., & Green, D. R. Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, Calif.: California Test Bureau/McGraw-Hill, 1977 (in press).

Wargo, M. J., Tallmadge, G. K., Michaels, D. D., Lipe, D., & Morris, S. J. ESEA Title I: A reanalysis and synthesis of evaluation data from fiscal year 1965 through 1970. Palo Alto, Calif.: American Institutes for Research, 1972.

Werts, C. E., & Linn, R. L. A general linear model for studying growth. Psychological Bulletin, 1970, 73, 17-22.

Williams, J. P. Learning to read: A review of theories and models. Reading Research Quarterly, 1973, 8(2), 121-146.

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

Wood, C. T., Cannara, A. B., Fagan, B. M., & Tallmadge, G. K. *Further documentation of state ESEA Title I reporting models and their technical assistance requirements, Phase I (Part Two)*. Mountain View, Calif.: RMC Research Corporation, August 1976.

Zimiles, H. Has evaluation failed compensatory education? In J. Helmuth (Ed.), *Compensatory education: A national debate. Volume 3, The disadvantaged child*. New York: Brunner/Mazel; 1970.

INDEX

Index, continued

Index, continued