DOCUMENT RESUME

ED 209 347 TM 810 902

AUTHOR	Haney, Walt
TITLE	Short-Term Impact Evaluation of Early Childhood Title I Programs.
INSTITUTION	Department of Education, Washington, D.C.
PUB DATE	Dec 80
NOTE	66p.
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Criterion Referenced Tests; *Early Childhood Education; *Evaluation Methods; Program Effectiveness; *Program Evaluation; Program Improvement
IDENTIFIERS	Aggregation (Data); *Elementary Secondary Education Act Title I; Evaluation Problems; *Impact Evaluation Model; RMC Models

ABSTRACT

This booklet briefly describes the potentials and problems of different ways of assessing the short-term impact of early childhood Title I (ECT-I) programs. It is intended for education officials and evaluators who already know something of Title I evaluation, but who may not be familiar with the special issues involved in evaluation of early childhood educational programs. Several issues in impact evaluation of educational programs in general are outlined, and the special problems associated with evaluation of ECT-I programs as distinct from Title I programs in later grades are addressed. Chapters III-VI describe four different approaches to evaluating short-term program impact, including the three models proposed by the United States Education Department for evaluating Title I programs in later grades. Chapter VII describes special issues relevant to criterion-referenced assessment of program impact. Chapter VIII discusses the key problems to be faced by anyone wishing to aggregate results of short-term impact evaluations of ECT-I programs across several cases that may or may not use the same approach to assessing impact. Chapter IX provides a summary, conclusions, and guidelines on the appropriateness of different kinds of short-term impact evaluations of early childhood Title I projects under different conditions. (Author/GK)

Short Term Impact Evaluations of Early Childhood Title I Programs

DEPARTMENT OF EDUCATION · UNITED STATES OF AMERICA

U.S. Department of Education

SHORT-TERM IMPACT EVALUATIONS OF EARLY CHILDHOOD TITLE I PROGRAMS

December 1980

Walt Haney
The Huron Institute
123 Mount Auburn Street
Cambridge, Massachusetts 02138

FOREWORD

This booklet has been prepared as part of a project sponsored by the United States Education Department (USED) on evaluation in early childhood Title I (ECT-I) programs. It is one of a series of resource books developed in response to concerns expressed by state and local personnel about early childhood Title I programs. The series describes an array of diverse evaluation activities and outlines how each of these might contribute to improving local programs. The series revolves around a set of questions:

- Who will use the evaluation results?
- What kinds of information are users likely to find most helpful?
- In what ways might this information aid in program improvement?
- Are the potential benefits substantial enough to justify the cost and effort of evaluation?

Together, the resource books address a range of issues relevant to the evaluation of early childhood programs for educationally disadvantaged children. The series comprises the following volumes:

- Evaluating Title I Early Childhood Programs: An Overview
- Assessment in Early Childhood Education
- Short-Term Impact Evaluation of Early Childhood Title I Programs
- An Introduction to the Value-Added Model and Its Use in Short-Term Impact Assessment
- Evaluation Approaches: A Focus on Improving Early Childhood Title I Programs
- Longitudinal Evaluation Systems for Early Childhood Title I Programs
- Evaluating Title I Parent Education Programs

The development of this series follows extensive field work on ECT-I programs (Yurchak & Bryk, 1979). In the course of that research, we identified a number of concerns that SEA and LEA officials had about ECT-I programs, and the kinds of information that might be helpful in addressing them. Each resource book in the series thus deals with a specific concern or set of concerns. The books and the evaluation approaches they describe do not, however, constitute a comprehensive evaluation system to be uniformly applied by all. Our feasibility analysis (Bryk, Apling, & Mathews, 1978) indicated that such a system could not efficiently respond to the specific issues of interest in any single district at any given time. Rather, LEA personnel might wish to draw upon one or more of the approaches we describe, tailoring their effort to fit the particular problem confronting them.

Finally, the resource books are not comprehensive technical manuals. Their purpose is to help local school personnel identify issues that might merit further examination and to guide the choice of suitable evaluation strategies to address those issues. Additional information and assistance in using the various evaluation strategies are available in the more technical publications cited at the end of each volume, and from the Technical Assistance Centers in the ten national regions.

TABLE OF CONTENTS

I. INTRODUCTION

The goal of a short-term impact evaluation is to estimate the immediate effect of a program on participants. This booklet briefly describes the potentials and problems of different ways of assessing the short-term impact of early childhood Title I (ECT-I) programs. It is intended for education officials and evaluators who already know something of Title I evaluation, but who may not be familiar with the special issues involved in evaluation of early childhood educational programs.

The book first outlines several issues in impact evaluation of educational programs in general, and then addresses the special problems associated with evaluation of ECT-I programs as distinct from Title I programs in later grades. Next, the middle section of the book (Chapters III-VI) describes four different approaches to evaluating short-term program impact, including the three models proposed by the United States Education Department (ED) for evaluating Title I programs in later grades (i.e., grades 2-12). Chapter VII describes special issues relevant to criterion-referenced assessment of program impact. Chapter VIII discusses the key problems to be faced by anyone wishing to aggregate results of short-term impact evaluations of ECT-I programs across several cases that may or may not use the same approach to assessing impact. Chapter IX provides a summary and conclusions together with some guidelines on the appropriateness of different kinds of short-term impact evaluations of early childhood Title I projects under different conditions. Although the aim of the book is largely introductory, references to sources of detailed information are provided.

Before we launch into the main issues, several explanations are needed.

- Early childhood Title I programs mean programs funded under Title I of the Elementary and Secondary Education Act of 1965 (as amended) which serve children in preschool (or prekindergarten), kindergarten; and first grade.*

- Impact evaluation refers to evaluations that aim at estimating the effects or impact of a program.

- Short-term refers to assessment at or near the end of the program.

Since Title I programs generally extend for one school year, this means assessing program effects around the end of the school year. Short-term evaluation can be contrasted with long-term or longitudinal evaluation, which attempts to estimate the effects of a program some time after its participants have actually left the program. A separate resource book deals with longitudinal evaluation of ECT-I programs (Kennedy, 1980).

The distinction between short-term and long-term evaluation is much more than an academic issue. It is easy to forget that not all important program goals can be addressed in short-term impact evaluations: some educational goals are not short-term. Early childhood programs, for instance, often aim not just at preparing children for the second grade, but also at helping them to become active learners and better citizens later in school and in life. In fact, the distinction is especially important for early childhood programs, because relatively modest intervention during what is often seen as a "critical period" can have long-lasting consequences. Some researchers have recently presented evidence of the long-term effects of preschool programs, effects that might not have been predicted on the basis of short-term evaluations (Lazar et al., 1977, Darlington, 1980).

* This definition of early childhood, which differs from those used elsewhere, is used throughout the project under which the resource books are being produced.

On a much smaller scale, Pedersen and Faucher (1978) in a long-term follow-up study found that important effects of one first-grade teacher did not become apparent until long after children had left her classroom.

The point of these examples is clear: short-term impact evaluation can address only the short-term goals of educational programs. If we assume certain connections between short-term learning and later achievement, short-term impact evaluation may help answer long-term questions; but it is important to recognize at the outset that it cannot directly address some long-term goals of ECT-I programs.

II. GENERAL ISSUES IN SHORT-TERM IMPACT EVALUATION

The key question to be asked before undertaking any evaluation--be it short-term impact evaluation or any other--is, why do it? This question is crucial because the special feature of evaluation, the one that distinguishes it from research, is that it is designed to provide information for decision making and program improvement. Hence the first issue to be addressed is how will short-term impact evaluation results be used? The second issue is, why give short-term impact evaluations of ECT-I programs special attention, apart from Title I programs for later grades? The third issue is, what characteristics make an impact evaluation technically sound? To examine the last issue fully, we must consider another general question: how should a short-term impact evaluation of ECT-I programs be designed? After discussing these general questions in this chapter, we turn in subsequent ones to the particular features of different short-term impact evaluation designs and the special programs likely to arise in applying them to ECT-I programs.

WHY CONDUCT A SHORT-TERM IMPACT EVALUATION?

Short-term impact evaluations are designed to provide information on the immediate effects of programs. In other words, such evaluations seek to answer the question of how children at the end of a program have changed as a result of participating in it. If we take the notion of evaluation seriously--that is, if we assume that it will contribute to better decision making to improve programs--then we must ask at the outset what kinds of decisions and what uses this sort of information can serve. If one does not carefully consider what information might be useful, then it is fairly likely that, however competently the evaluation is carried out from a technical point of view, its results will not be used.

There are several broad classes of use for Title I evaluation. In our previous review of ECT-I programs we found that reported uses of evaluation information differ substantially at the federal, state, and local levels (Bryk, Apling, & Mathews, 1978, Chapters 5, 7, and 8). Since this resource book is intended mainly for practitioners at the local level--and indeed, since this is the level at which education, as opposed to simply educational administration, takes place--let us focus on the potential local uses of evaluation. At this level, eight types of evaluation use were reported:

- Meeting state reporting requirements
- Assessing program effectiveness
- Improving programs
- Needs assessment
- Selection of students
- Staffing decisions
- Pupil diagnosis
- Promoting and assessing parent involvement.

Some of these uses, of course, can overlap. Assessing program effectiveness obviously can contribute to program improvement. But even this rough listing makes it clear that short-term impact evaluation is relevant for only some uses. It is not, for example, directly relevant to decisions about individual students--though information used to evaluate program impact often can also be used in other ways to help make decisions about individuals. Short-term impact evaluation is, however, potentially relevant to decisions about programs, for meeting reporting requirements and for assessing program effectiveness. The ED has not mandated short-term evaluations for ECT-I programs as it has for later-grade Title I programs--for reasons we will discuss in a moment. What this means, however, is that short-term impact evaluations of ECT-I programs should not be viewed simply as a means of meeting federal reporting requirements. Instead one ought to consider the potential utility of ECT-I short-term impact evaluation for other purposes; for example for making decisions about programs, improving programs and assessing program effectiveness. In the abstract, the potential utility of any particular ECT-I program cannot be determined. Nevertheless, before any technical or design issues are considered, the first question one ought to ask is whether a short-term impact evaluation is likely to be useful, and if so, to whom and in what context. In other words, is short-term impact evaluation of an ECT-I program likely to produce information sufficiently useful to justify its cost? If the answer is no, then proceed no further; the evaluation should not be done. If the answer is yes, then one must next consider how to conduct the evaluation.

SPECIAL ISSUES INVOLVED IN IMPACT EVALUATION OF EARLY CHILDHOOD PROGRAMS

What makes impact evaluations of early childhood programs especially difficult? Why does ED treat the impact evaluation of ECT-I programs differently from that of later-grade Title I programs?

The ED system for evaluating the impact of later-grade Title I programs consists of three basic models: a norm-referenced group design; a comparison group design; and a regression design. The conditions under which one of these evaluation models may be appropriate to some ECT-I programs will be discussed in Chapters III, IV, and V. Here, let us recount the reasons why ED has not simply mandated the application of the evaluation models to ECT-I programs. Three main considerations limit the potential usefulness of the models for ECT-I programs:

- First: The models were developed mainly to assess the impact of later-grade programs on reading, math and language arts. This range of program goals is too narrow for many early childhood programs whose objectives are broader, for example often including social, emotional, and psychomotor development.

- Second: Standardized tests such as are often used in later-grade Title I program evaluations raise special problems when used with young children. For example, standardized tests are not available for some common ECT-I program goals, and even when available, often have inadequate norms and relatively low reliability.

- Third: Early childhood programs often have long-range goals that simply cannot be encompassed in short-term impact evaluations.

For these reasons, ED decided not to employ the same system for evaluating ECT-I programs as has been developed for later-grade programs. Nevertheless, short-term impact evaluation may still be feasible and desirable for certain ECT-I programs. The purpose of this resource book is to describe alternative ways of evaluating the short-term impact of ECT-I programs, and the conditions under which they may be applied.

TITLE I TECHNICAL STANDARDS

While ED has not mandated impact evaluation for ECT-I projects in terms of any specific evaluation models, it has set forth technical standards relevant to any evaluation of Title I project effectiveness or impact. These deal with (1) valid assessment of program goals; (2) representativeness of evaluation findings; (3) reliability and validity of evaluation instruments and procedures; and (4) evaluation procedures that minimize error. How these four issues relate to Title I programs in grades 2-12 has been described in ED's evaluation regulations for Title I (Federal Register, October 12, 1979, subpart F, section 116a.50). Here, let us describe them only briefly as they apply to ECT-I programs.

Valid Assessment of Program Goals

Whatever the goals of an ECT-I program, an impact study, if it is to be useful, should be based on a valid assessment of those goals. Evaluation need not necessarily encompass all goals of the program, but however thoroughly it addresses those goals, the impact evaluation should comprise a valid assessment of at least some significant program goals. If an impact evaluation addresses only a subset of a program's goals, the goals not encompassed should be clearly identified, and the evaluator or evaluation report should make it completely clear that the impact evaluation does not constitute overall evaluation of the program's worth.*

Representativeness of Evaluation Findings

The evaluation should be conducted so that the conclusions drawn apply to the persons (children or their parents), schools, and agencies served by the ECT-I program concerned. This means that it should include all, or a representative sample, of the persons, schools, or agencies the program serves.

* Strictly speaking, of course, even if an evaluation thoroughly addresses all program goals, it cannot be said to constitute an overall assessment of program worth. Programs may, for example, have unintended or side effects not covered in any stated program goals. For this reason, some evaluators have advocated goal-free evaluation, that is evaluation aimed at assessing both intended and unintended effects of programs. Without getting into the general debate over this question, let us observe that some advocates of goal-free evaluation (e.g., Scriven, 1974) have suggested that it can best be carried out by external evaluators not directly connected with the program they are evaluating.

Reliability and Validity of Evaluation Instruments and Procedures

Instruments used in impact evaluation must consistently and accurately measure the attainment of project objectives. They must be appropriate in terms of such factors as ages and backgrounds of the persons served by the project. For example, using separate test answer sheets is generally not appropriate for the preschool to grade-1 age range, since children of these ages are usually not yet able to record answers accurately on separate answer sheets like those used in assessments of older children. (See the resource book Assessment in Early Childhood Education by Haney & Gelberg, 1980, for more information on this and related issues.)

Evaluation Procedures That Minimize Error

Error should be minimized by proper administration of evaluation instruments, quality control procedures that ensure accurate scoring and transcription of results, and choice of analysis procedures whose assumptions apply to the data obtained from the evaluation. In ECT-I programs, the proper administration of evaluation instruments is especially important, since young children's performance is influenced more than that of older children by variations in instructions and practice preceding the administration of evaluation instruments. Since individually administered assessments often are more appropriate for young children than group administered assessment, attention needs to be given to whether or not errors may be introduced through unwarranted variations in administration procedures.

DESIGNING A SHORT-TERM IMPACT EVALUATION

The four considerations described above apply to any impact evaluation, short-term or otherwise. Beyond such general issues of technical quality, one must also consider the appropriate design for a short-term impact evaluation of ECT-I programs.

A short-term impact evaluation aims at estimating the immediate effect of a program on participating children. The program effect may be defined as the difference between the status of participating children at the end of the program and the status they would have attained had they not received ECT-I services. This is often expressed as "program effect equals observed status at end of program minus status expected without program." Designs for short-term impact evaluations differ mainly in how they estimate the status children would have attained had they not received special services. In chapters III-VII we discuss five designs for estimating program impact or effect:

- Norm-referenced approach
- Comparison group approach
- Regression approach
- Value-added approach
- Criterion-referenced approach.

The first three approaches correspond to models A, B, and C in ED's system for evaluation of Title I programs in grades 2-12 (Tallmadge & Wood, 1978, 1980). The value-added approach is a method of using children's ages to estimate their expected no-treatment status. The criterion-referenced approach treated separately in Chapter VII uses explicit program objectives as a basis for estimating impact.

Before describing each of these five approaches in detail, let us briefly describe a strategy for deciding which approaches to consider, as depicted in Figure 1. First, one must consider whether the results of a short-term impact evaluation are likely to be sufficiently useful to justify its costs. There is of course no clear way to determine this precisely. No one has ever attempted to develop a system for analyzing the cost effectiveness of evaluation. Nevertheless, before forging ahead with a short-term impact evaluation, this issue should at least be considered informally. If one decides to go ahead, the next question is whether the same outcomes are to be assessed for all participating students. If not, that is if different outcomes or objectives are to be assessed for different children, then a criterion-referenced approach to impact assessment should be considered.* If common outcomes are to be assessed for all children, then the next question is whether an appropriate comparison group of children is available. If not, one should consider using the norm-referenced or value-added approach. If an appropriate comparison group is available, one can consider using either the comparison group or the regression approach to evaluating short-term impact.

This strategy for deciding which approaches to consider is very general. Different factors bear on these considerations at the prekindergarten, kindergarten, and first grade levels, as we will explain in subsequent chapters. Nevertheless, as an initial guide for thinking about which approaches to consider, this strategy may prove useful.

* Note that this question does not necessarily pertain to whether the program to be evaluated is an individualized one. Individualized approaches may after all employ different methods for promoting the same outcomes for all children in a program.

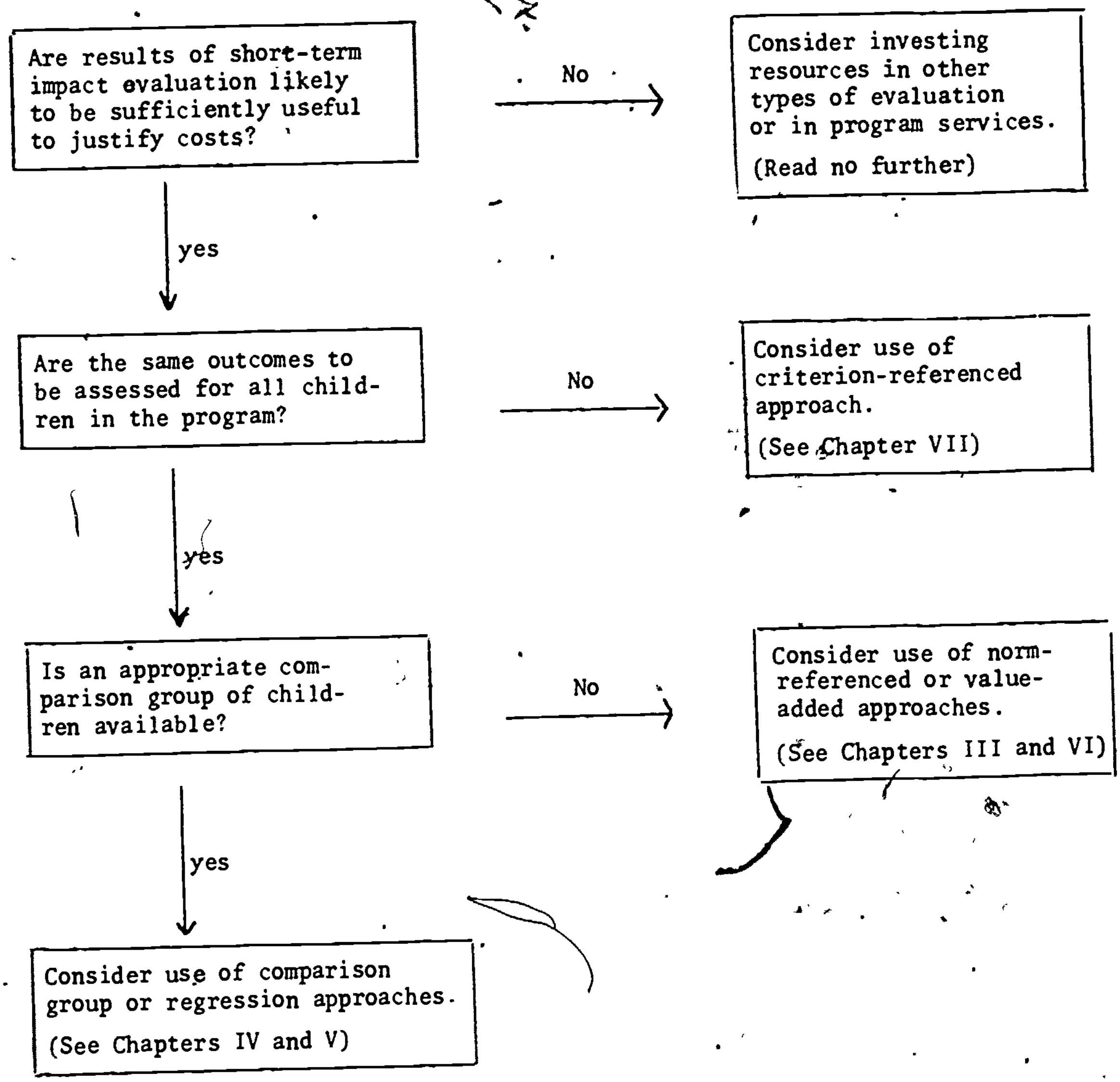
Are results of short-term impact evaluation likely to be sufficiently useful to justify costs?
No
Consider investing resources in other types of evaluation or in program services. (Read no further)
yes
Are the same outcomes to be assessed for all children in the program?
No
Consider use of criterion-referenced approach. (See Chapter VII)
yes
Is an appropriate comparison group of children available?
No
Consider use of norm-referenced or value-added approaches. (See Chapters III and VI)
yes
Consider use of comparison group or regression approaches. (See Chapters IV and V)

Figure 1. A Strategy for Deciding Which Approaches to ECT-I Short-Term Impact Evaluation to Consider.

Before we describe each of these approaches, it may be helpful to explain why we treat the criterion-referenced approach separately with respect to ECT-I programs; whereas, with respect to later-grade programs, it is considered as a variant of Model A, B, or C (the norm-referenced, comparison group, or regression approach, respectively). In later-grade programs, ED intends to aggregate effects estimates across different Title I projects to obtain overall estimates of effectiveness of Title I programs in reading, math, and language arts. However, because of the wider diversity of ECT-I program goals and the special problems in early childhood testing and evaluation, aggregation across all ECT-I programs is less feasible. Hence, in the case of ECT-I programs, we are free to consider the criterion-referenced approach in its own right, without dealing directly with the problems of equating its results with those of the other approaches. Nevertheless, since aggregation of impact evaluation results across some types of ECT-I programs may be both feasible and of interest, Chapter VIII provides a discussion of the conditions under which it may be possible to aggregate results across the different approaches.

Each of the following five chapters does three things:

- First, it describes one of the five general approaches to designing short-term impact evaluations.
- Second, it discusses the strengths and weaknesses of that approach with respect to ECT-I programs in general.
- Third, it summarizes the likely utility of the design, at the prekindergarten, kindergarten, and first-grade levels.

III. NORM-REFERENCED APPROACH

Model A, or the norm-referenced approach, is the most commonly used of the three ED models for evaluating later-grades programs (Anderson et al., 1978). Partly, this is because it seems the easiest to implement: it does not require gathering data on a comparison group. On its surface the model may seem easy to use with ECT-I programs as well; but it should be stressed that this approach poses special difficulties with respect to early childhood programs.

The norm-referenced approach is based on assessment of children at both the start and the end of a Title I program--commonly referred to as pre- and posttesting--and then comparing their actual performance with expected performance derived from norms tables available for the assessment instrument. Typically, the instrument used in this approach is a nationally normed standardized achievement test. The essential assumption in this approach is that without Title I services, children would maintain their status relative to the norm group from the start to the end of the program. Hence, the norm-referenced status of children at the start of the program is used as an estimate of what their status would have been at the end of the program had they not received Title I services (the no-treatment expectation). Program impact or effect is calculated by subtracting children's norm-referenced status at the start of the program (pretest performance in norm-referenced terms) from their norm-referenced status at the end of the program (posttest performance in norm-referenced terms).

The norm-referenced scale most commonly used in this model is the percentile score, although others such as stanines, deciles, or standard deviations also could be employed. In proposing rules and regulations for later-grade Title I evaluation, ED described this approach using percentile scores as follows:

> The amount of gain attributable to Title I is computed [using appropriate statistical procedures] by determining the pretest percentile status of the Title I children from the norms table [for the test] and obtaining the expected performance from the posttest norms table, assuming the Title I children would maintain the same percentile status. The Title I impact is the difference between the observed posttest score and the expected performance.
>
> (Financial Assistance. . . , February 7, 1979, p. 7916)

In discussing the rules for implementing the norm-referenced approach, the 1980 User's Guide suggested four requirements:

- The model requires using normative data to establish the no-project [or no treatment] expectation.
- The test level should match the functional level of the students. ... and should contain items that reflect the instructional content of the project.
- The choice of which test form to use at pretest and posttest should be determined by the design used by the test publisher in the development of the norms.
- All testing must be accomplished no earlier than two weeks before and no later than two weeks after the midpoint of the period during which the normative data were collected unless the norms are linearly interpolated or extrapolated (Tallmadge & Wood, 1980, pp. 39-40).

These requirements suggest why it is difficult to evaluate the short-term impact of ECT-I programs using the norm-referenced approach. First, many early childhood tests and instruments do not have adequate norms. In fact, there are few, if any, adequately normed tests for some of the common goal areas of ECT-I programs, such as psychomotor and social development. Instruments measuring attitudes toward schooling and emotional attributes of children, for example, often lack norms. A related problem is that normative interpretations of early childhood test results are strongly dependent on children's prior educational and social experience. For this reason differential norms are available for some tests used with young children. One commonly used standardized test series, for example, offers two sets of norms for the beginning of first grade--one for children who attended kindergarten and another for those who did not. The same raw scores for the alphabet subtest of this instrument, when interpreted in terms of these two sets of norms, vary by as much as 40 percentile points.

The problem here is not that a single instrument can have different sets of norms. Many instruments have two or more sets of norms derived from different norming samples, and hence relevant to different populations. Rather the problem is that young children typically have varied sorts of preschool and early school experience, and early childhood test norms, even when available, rarely control adequately for diversity of early experience, or even identify directly the previous school or preschool experience of the norming sample.

Normative interpretations of early childhood test results vary widely depending on children's previous educational experience, because young children develop rapidly. When first given instruction they tend to learn certain basic skills like letter and number recognition quite quickly. This relates to another issue in using the norm-referenced approach with ECT-I programs. The rapid development of young children is also why early childhood tests tend to cover a relatively wide range of skill levels even though they may be designed for only a single grade level. In other words, the grade span coverage of individual test levels tends to be narrower in the early grades than in later grades. The A level of 1978 Gates-MacGinitie Reading Test, for example, is intended to be appropriate, according to its publisher, for only the latter part of first grade. For more information on the special issues of early childhood testing and instrumentation see Assessment in Early Childhood Education (Haney & Gelberg, 1980).

In summary, the norm-referenced approach to assessing short-term program impact cannot be generally recommended for all ECT-I programs. It is impossible to use when normed instruments whose content matches the goals of the ECT-I program are unavailable, and in general may be more difficult to apply appropriately with programs serving younger children, for example at the prekindergarten and kindergarten levels. At the first-grade level the norm-referenced approach may prove more feasible, for example in evaluating the short-term impact of a program which aims at developing early reading skills for which a norm-referenced instrument is available. Nevertheless, even at this level the norm-referenced approach should be used with caution. In particular, one needs to consider whether norm-referenced results may reflect shifts in content of pre- and posttests or significant differences in previous school experience of program children and norm group samples. Also, it should be noted that norm-referenced estimates of ECT-I program impact should not be compared directly with similar results for later-grade programs. At the early childhood level impact estimates may be larger simply because of the fact that young children develop more rapidly than older children (see Notes for further information on this point).

IV. COMPARISON GROUP APPROACH

The comparison group approach involves assessment of both the group of children receiving Title I services and a group of children essentially comparable to the Title I children in all respects other than not receiving Title I services. The status of the comparison or control group children serves as an estimate of what Title I children would have been like had they not received Title I services. This approach corresponds to Model B in ED's system for evaluating Title I programs.*

As ED's User's Guide points out, "Model B, if implemented correctly, can be the most rigorous of the models because local students who are similar to Title I students provide the most accurate no project expectations" (Tallmadge & Wood, 1980, p. 55). Indeed, this approach derived from the randomized control group model of experimental research. From the scientific point of view this model is ideal, because random assignment of individuals to treatment and control groups helps to guarantee that the two groups differ only in that the former group receives treatment (corresponding to Title I services in the present discussion). Random assignment requires that all individuals who are to be assigned to these groups be identified in advance. Once they have been clearly identified and randomly assigned to treatment and control groups (analogous to project and no-project groups), it is possible to calculate precise mathematical probabilities that the two groups will differ from each other in pertinent ways.

* This approach has often been called the control group approach. However, since official ED rules and regulations refer to it as the comparison model, we use this terminology (Financial Assistance. . .', October 12, 1979).

In many real-life situations it is simply not possible, or even desirable, to use this approach. In Title I, for example, randomly assigning children to programs clearly is contrary to the program goal, which is to serve the most educationally disadvantaged children. For this reason, ED in its evaluation system for Title I does not advocate use of the randomized model. Instead, in describing Model B it has chosen to recommend the more flexible comparison group approach: first selecting Title I children according to appropriate criteria, and then locating a control group as much like the Title treatment group as possible. The User's Guide states that "the process of selecting a comparison group is not particularly hazardous--as long as the two groups are sufficiently similar" (Tallmadge & Wood, 1980, pp. 55-56). The Guide recommends that "comparison-group students are most likely to be found in the non-Title-I school in the district that just missed qualifying for Title I services" (p. 57), and suggests:

> Select the comparison group students by locating the non-Title I school (or schools) in the district (or a nearby district) that is most like the school serving the Title I students. Identify students for the comparison group in the non-Title I school by using the same objective measure(s) as was used to identify the project students. In the case where pretest scores are used for selection, do not apply the same pretest cutoff score to select the two groups. Instead, determine the percentage of Title I students in the Title I school and select the same percentage of low-scoring students in the non-Title I school for the comparison group.
> (Tallmadge & Wood, 1980, p. 58)

The User's Guide suggests that the posttest score of the comparison group can be used as the no-treatment expectation only if the mean pretest scores of the two groups differ by 1 NCE or less.* If the mean pretest scores of

* The NCE or normal curve equivalent is a metric, similar to the stanine, used to interpret test scores. Approximately 11 NCEs are equal to one stanine. For an explanation of the NCE see Tallmadge & Wood, 1978.

the treatment and comparison groups differ by more than this amount, it is recommended that some type of adjustment must be made to compensate for the initial pretest differences (see Tallmadge & Wood, 1980, p. 59-60). However, if the pretest scores of the treatment and comparison groups differ by a substantial amount, say more than 4 NCEs, then one should question the appropriateness of the comparison group model, even with adjustment for initial differences. In other words, if treatment and comparison groups differ by more than a little (i.e., more than 1 NCE) but less than a moderate amount (less than 4 NCEs), then the pretest scores of the two groups can be used in statistical calculations to estimate what the status of the treatment group would have been had it not received the treatment.

These suggestions are, however, only rough rules of thumb. If the pretest scores of treatment and control groups are <u>similar</u> but for <u>different reasons</u>, then these guidelines may be misleading. One example from a large-scale early childhood evaluation will help illustrate this problem. The national evaluation of the Follow Through (FT) program was based primarily on the comparison group approach embodied in Model B of the Title I evaluation system. An attempt was made to select comparison groups from neighboring schools which were essentially similar to the groups of children receiving FT services in each FT project. In many projects the pretest scores of FT children and comparison group children were very similar. However, in some cases a higher proportion of FT children than of comparison-group children had Head Start experience, and this apparently accounted for the similarity in pretest scores for the two groups. A special analysis of the FT data, taking previous Head Start experience into account, showed that some evaluation results could change significantly when this differential in preschool experience was controlled. (See Haney, 1977, for an account of how control groups were selected for the FT evaluation, and Weisberg & Haney, 1978, for a description of how controlling for differential preschool experience for treatment and comparison groups could change evaluation results for some projects.)

This example illustrates the main difficulty in applying the comparison group approach in evaluating ECT-I programs--namely, that of finding altogether appropriate comparison groups. This problem appears to be the main reason for the relatively rare use of Model B in evaluating later-grade Title I programs (Anderson, et al., 1978). It can be especially severe for ECT-I programs. Children are selected for ECT-I programs in a diversity of ways (Yurchak & Bryk, 1978), some of which may be impossible to duplicate in finding comparison groups. For later-grade programs, there often is available a population of children already enrolled in school from which comparison group children can be selected in ways similar to those used in selecting Title I children. For many prekindergarten and kindergarten ECT-I programs, however, often there is no source of control group children easily available. Thus selection of comparison group children can be very expensive if not altogether impossible. Even if ECT-I comparison groups could be found whose pretest scores are essentially similar to those of ECT-I children, the similarity may be due to differential preschool experience. In such cases it is necessary to employ statistical controls for preschool experience in deriving a no-treatment expectation for the ECT-I program from the comparison group posttest scores.

In summary, the comparison group approach is potentially the strongest strategy for assessing short-term impact of ECT-I programs. If a comparison group can be located which is essentially the same as the group of ECT-I participants then it can provide a clear indication of what ECT-I children would be like if they had not received ECT-I services. The main problem with this approach is that it often is hard to locate a group of comparison children who are in fact "essentially comparable" to the children selected to receive ECT-I services. Even if average pretest scores of the Title I and comparison groups are nearly equal, it may be because they differ in other important respects. At the early childhood level, the comparison group approach likely has broader applications for first grade programs (where comparison groups may be available in neighboring schools) than for prekindergarten programs (for which preschool age children are specially recruited, and for whom there simply may be no easily available comparison).

V. REGRESSION APPROACH*

A third approach to estimating short-term program impact also employs a comparison or control group, but does so in a different way than the simple comparison group approach. In the regression approach, the comparison and treatment groups are not assumed to be essentially equivalent at the start of the program. Instead, the differences between the two groups are explicitly controlled in the process of assigning individuals to each. A decision rule is established for assignment to treatment and control groups. Then, since the exact basis for assignment to each of the groups is known, statistical analysis of these groups at the start and end of the treatment can be used to derive an estimate of what the treatment group would have been like had it not received the treatment.

This approach corresponds to Model C in ED's system for evaluating Title I programs in grades 2-12. The Model C regression approach has been described by ED as follows:

> In the ... Regression Model, a group of children is divided into Title I and comparison groups based on a pretest cutoff score. Title I services are provided to children scoring below the cutoff. Children scoring above the cutoff are the comparison group for the evaluation. Expected performance is estimated from the pretest and posttest scores of the comparison group by applying a statistical procedure known as the regression model. This model, when properly applied, will yield an estimate of expected performance that takes into account differences between the two groups that are not the result of Title services.
> (Financial Assistance. . . , February 7, 1979, p. 7916).

*. This approach has often been called the special regression model, since other approaches--for example, the control group approach with statistical adjustment and the value-added approach--may use regression analysis. However, since official ED regulations term it simply the regression model, we omit the word "special."

Figures 2 and 3 illustrate the regression model. Figure 2 shows a simple scatterplot of pretest and posttest scores. Such a plot might easily be drawn for any set of pre- and posttest scores, such as fall and spring scores for children in any school. Note that children with the same pretest scores may have different posttest scores, but that in general those with higher pretest scores tend to have higher posttest scores. One way of representing this tendency is with a regression line, also shown in Figure 2. Although the mechanics of calculating regression lines can get very complex, the basic idea is really quite simple--exactly the same as in graphing linear equations, which is taught in high school algebra courses. The regression line shown in Figure 2 is just such a linear equation:

$$Y(\text{posttest score}) = 10 + .8X(\text{pretest score})$$

The regression model employs such regression equations to derive no-program (or no-treatment) expectations for children who do in fact receive the program (in this case Title I services). First, children whose pretest scores are below a particular level (the cutoff point) are assigned to receive Title I services, and those whose pretest scores are above that level are assigned to the comparison group. How such regression lines might look is shown in Figure 3. The downward projection of the regression line of the comparison group is then used to estimate what the status of program children would have been had they not received the treatment. Subtracting these "no-treatment" expectations from the actual scores of the program group yields an estimate of program impact. In the example shown in Figure 3, this estimate--the difference in height between the two regression lines--is 5 points.

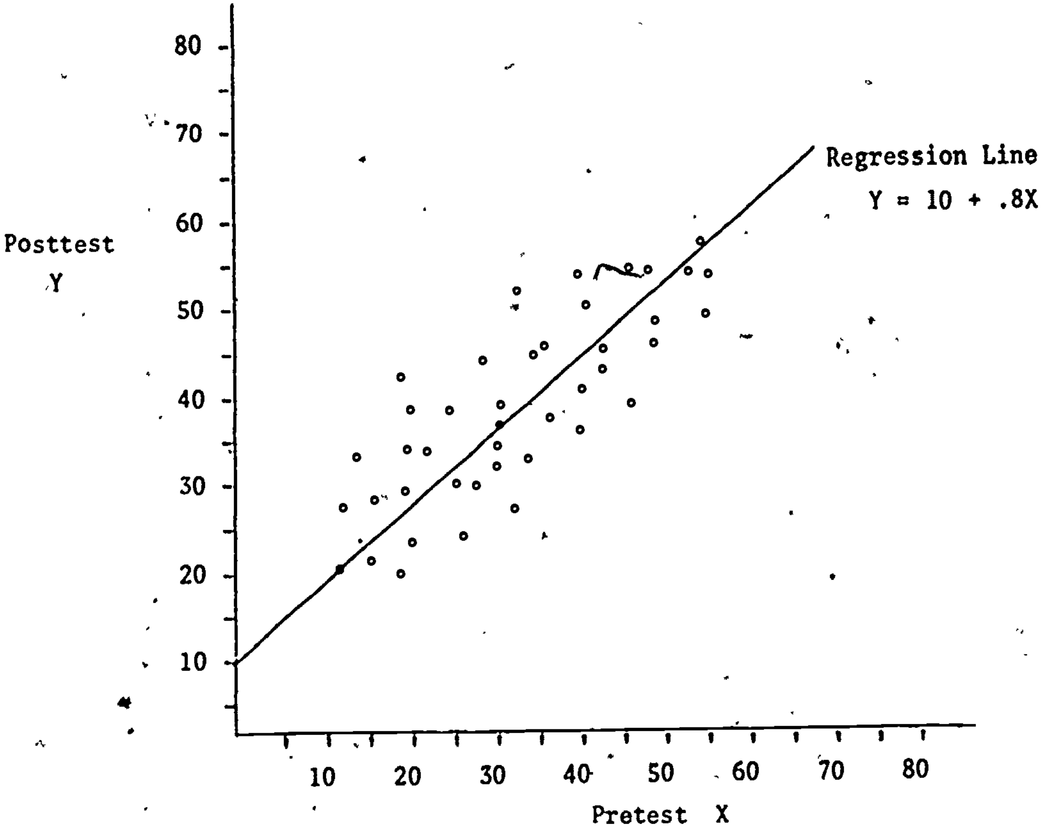

Figure 2: Regression Line of Posttest and'Pretest

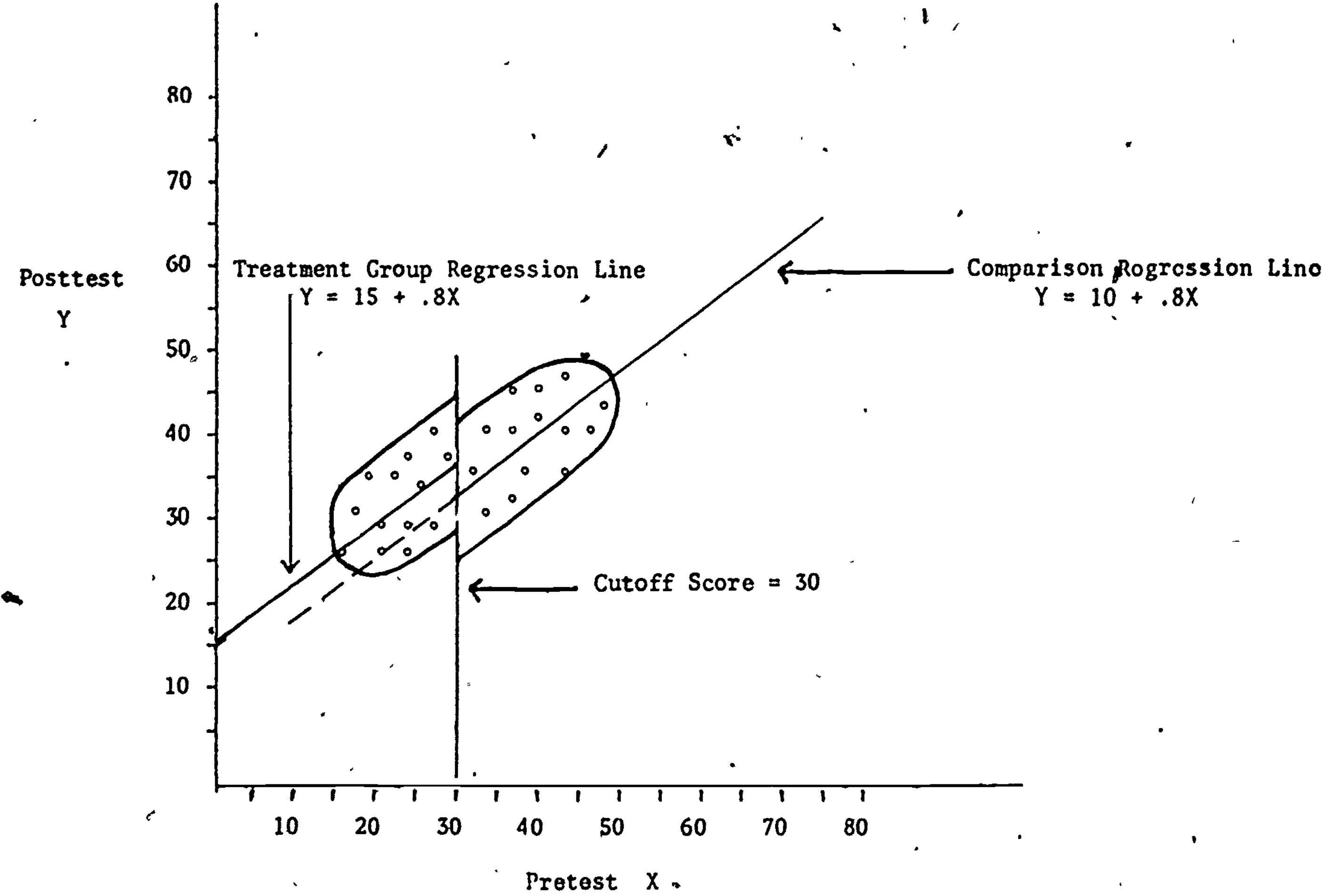

Figure 3: <u>Regression Lines Calculated Separately Under Special Regression Approach for Treatment and Comparison Groups</u>

Measuring program impact in this way is moderately complex computationally and also can be fairly expensive, because like the comparison group approach, it requires the tbsting of students not in the program. The ED User's Guide suggests that Model C should probably never be implemented with fewer than thirty students in each group (Tallmadge & Wood, 1980, p. 71). The approach places few constraints on what tests are administered at the start and end of the program. All that is necessary is at least a moderate correlation between pre- and posttest. The User's Guide suggests that the model should not be implemented if the correlation between the pretest and posttest measures is less than 0.6 for the total group or 0.4 for the comparison group (p. 22). In fact, however, other things being equal, the absence of moderate correlation would merely transform the regression approach into a comparison group approach.*

One practical difficulty of this approach is that assignment of children to the program and comparison groups must be based strictly on the pretest score or on some composite (for example, a weighted combination of teacher judgment and pretest results). Another difficulty is that if results are to be unambiguously interpreted, then the regression lines calculated separately for the treatment and comparison groups must be parallel. If they are not, this suggests that there may have been problems in the instrumentation used, or that treatment and comparison groups differed in some important way that affected the relationship between pre and posttest scores. Nonparallel lines may also result if the treatment

* Of course, other things rarely will be equal, since in this approach treatment and comparison group children are selected on the basis of differences in pretest scores, and a range of other variables are likely related to such differences.

was differentially effective for different pupils (which is sometimes called a treatment-aptitude interaction; Cronbach & Snow, 1977). Differential treatment effectiveness may be entirely desirable from a program point of view; but from an evaluation standpoint, it may be hard to tell whether nonparallel regression lines are due to this factor or to other ones, concerning selection or instrumentation.

These constraints limit the usefulness of this approach to evaluating the short-term impact of ECT-I programs. First, given the diversity of recruitment and selection procedures used for ECT-I programs, and particularly prekindergarten programs (Yurchak & Bryk, 1978), it may be impossible to base program assignment strictly on a pretest or composite score. Second, since a comparison group is required, the problem encountered with the comparison group approach--that of locating comparison groups for ECT-I programs--also arises here. Children often are recruited in special ways for preschool Title I programs. Since there often is no clearly identified population from which they are selected, it may prove extremely difficult --if not impossible--to apply this approach to some ECT-I programs. Third, there is the familiar problem of early childhood testing and instrumentation. Because young children develop so rapidly, floor and ceiling effects may occur--either the pretest or the posttest may be too easy or too hard for either program group or comparison group children. This problem may be especially severe with respect to the regression approach. The required assignment of program group and comparison group children strictly on the basis of some pretest and composite, coupled with the fast rate of development of young children, may produce ceiling effects in comparison group children at posttest time.

In summary, the regression approach to estimating short-term impact of ECT-I programs has both strengths and weaknesses. One strength is flexibility. It is not necessary to administer the same test as a pre- and posttest. The major weakness is that this approach requires use of a comparison group and selection of participating children in a precisely specified manner (specifically, participants have to be selected on the basis of falling below a cutoff score on a pretest or on a composite of a pretest score and/or other information). This requirement suggests that the regression approach may be more feasible for use at the first grade and kindergarten levels. At those levels ECT-I participants are often selected from groups of children already enrolled in school. It is less feasible for prekindergarten programs (and some kindergarten ones) which use special recruitment procedures to locate participants among children who are not yet enrolled in school.

VI. VALUE-ADDED APPROACH

A fourth approach to estimating short-term program impact is called the value-added approach. It is based on a projection of individual students' status across the duration of the program. It uses information on children's pretest scores and ages, the time interval between pre- and posttests, and children's background characteristics to project explicitly the growth each child would have achieved without treatment. The actual performance of individuals at posttest time can then be compared with these projections to estimate program effect for each. Such individual program-effect estimates are sometimes referred to as the value added by the program; hence, the name given to this approach.*

The value-added approach, though less well known than the approaches described earlier, has been used in several early childhood evaluations (Smith, 1973; Weisberg, 1974) and has received some attention in the technical literature (Bryk & Weisberg, 1974, 1976, 1977). Since the approach is not widely known, let us first illustrate it with a simple example before describing its key assumptions and the possible problems in its use for short-term impact evaluation of ECT-I programs.

The value-added approach was applied by Bryk & Weisberg (1976) to data drawn from the national evaluation of the Head Start Planned Variation (HSPV) program. In that evaluation, a variety of tests were administered

* The name given to this approach should not be misunderstood. Other approaches to estimating program impact indicate "value added" by a program in the same sense that this approach does, but in different ways. For more information on this approach to measuring short-term impact see the Bryk & Woods (1980) booklet in this series and the references cited at the end of this booklet.

at both the start and the end of the HSPV program. Bryk & Weisberg, however, applied the value-added approach only to the Preschool Inventory, so that is the test discussed in this example. In order to evaluate the short-term impact of HSPV,* these investigators first determined the regression relationship between PSI pretest scores and children's ages (expressed in months). Via regression analysis they estimated that for each month of increase in age, a child scores an average .38 points higher on the PSI. Thus, a child who scored 18 on the PSI pretest at age 56 months could be expected, at posttest time at age 63 months, to score 18 + .38 (63-56), or 20.66. Subtracting this no-treatment expectation from the child's actual posttest score of 23, Bryk and Weisberg estimated that the program effect, or value added by the program, for this particular child was 23 - 20.66 = 2.34. Similar estimates for all children in the program could then be averaged to obtain an estimate of short-term program impact.

This is a very simplified example of the value-added approach. Bryk and Weisberg (1976) go on to illustrate a more elaborate application of the approach in which a no-treatment expectation is based not just on the child's age, but also on the child's race and sex, and mother's education. For details, including more elaborate applications, see the resource booklet on the value-added approach (Bryk & Woods, 1980).

The example above, though simple, serves to illustrate the essential features of this approach and thus will serve as a basis for our discussion of the conditions under which it may be applied to ECT-I programs.

* HSPV actually encompassed several different instructional programs, and effects of each of these were estimated separately. For the sake of this illustration, however, we will not elaborate on these differences, but will refer to HSPV as if it were a single program.

The main idea behind the value-added approach is to use information on the natural growth of children--in terms of their test scores or other characteristics--to predict what their status would be at posttest time if there had been no intervention. In this way the value-added approach does away with the need for using a comparison group as a basis for estimating no-treatment expectations.

The major technical problems in this approach stem from required assumptions about children's growth. The relation between pretest score and age and background variables is used to estimate the expected growth of each child between pre- and posttest. It is assumed that during that interval, individual growth increases steadily with age, and that the relationship between children's age and test scores at pretest can be used to estimate how they would have performed at posttest time had they not received treatment. The latter point refers to what is sometimes called the stable-universe assumption in the child development literature (see, for example, Kodlin & Thompson, 1958) and is a basic assumption in any attempt to draw longitudinal inferences (e.g., expected growth in the absence of ECT-I) from cross-sectional data. This assumption means simply that individual growth is independent of children's cohort or age group--for example, it assumes that absent any special intervention, children born in January 1978 will grow at the same rate as children born in June 1978.

Problems with this assumption can arise in several different ways. When ECT-I programs deal with relatively homogeneous groups, the assumption may be reasonable. But there may be historical trends causing children born at different times to differ. If ECT-I participants have different backgrounds--for example, different preschool experiences, as in the FT case cited above on page 21--then the assumption may not hold. Also, even if the stable universe assumption is valid for the population being studied, the process of selecting the program groups may introduce a problem. For example, the oldest children in an ECT-I preschool might be delayed entrants into a school group, and the youngest somewhat more precocious than average.

Another technical problem can arise when extrapolations beyond the observed data are required. When the value-added approach is used to predict the expected posttest scores of a group of children who will be considerably older at posttest time than at pretest, one implicitly assumes that the relation of age to test score apparent at pretest time will still be valid at posttest time. For example, if ECT-I participants are 45 to 57 months old at pretest time and the program lasts for nine months, in order to predict posttest scores the evaluator must extrapolate the model into the age range of 54 to 66 months. Such extrapolation--considerably beyond the originally observed age range--can raise real problems in the application of the value-added approach and is not in general to be recommended.*

This issue has implications for the short-term impact evaluation of ECT-I programs. First and foremost is the familiar problem of testing and instrumentation. Floor and ceiling effects can cause special difficulties in the value-added approach, because this approach depends upon the assumed

* Note that the ratio of the extrapolation range to the age range at pretest time is important in determining the efficiency of the value-added estimates. Strenio (1977) points out that as the ratio increases (i.e., a larger extrapolation relative to natural age variation), the precision of the estimation decreases.

stable relationship between age and test scores for all children (or, when controls are introduced for background variables, all children within single categories of variables which are controlled). The second potential difficulty is that the value-added approach is simply not appropriate for use in evaluating outcomes (including some test scores) that do not show natural increase with age across the duration of the program to be evaluated. By the first grade, for example, most children have developed gross motor skills involved in skipping or running, so changes in such gross motor skills would not show much, if any, relationship with age at the first grade level. Because the correlation between age and test scores, at least within grade level, tends to diminish as children get older, this approach will generally be less appropriate for older than for younger children.

In summary, the value-added approach to measuring short-term impact of ECT-I programs, like other ones, has both strengths and weaknesses. Its major strength is that it does not necessarily require a comparison group. Its major weaknesses are that (1) it is appropriate only for the assessment of skills or attributes which show a natural development with age over the duration of the program; (2) selection procedures may disguise the age-skill development relationship among a particular group of program participants (thus either precluding application of the value-added approach or necessitating reliance on some external comparison group as a source for deriving appropriate age-skill development projection); and (3) the value-added approach requires some reasonably complex statistical calculations. Since young children typically change and develop more rapidly than older ones, this suggests that the value-added approach may generally be more appropriate for ECT-I programs serving prekindergarten and kindergarten children than for those serving first-graders (see Notes for further information).

VII. CRITERION-REFERENCED APPROACHES

Criterion-referenced approaches to short-term impact evaluation compare performance at the end of the program with some specified criterion or standard of performance. As in the norm-referenced and value-added approaches, only the performance of program participants need be assessed. Criterion-referenced approaches to impact evaluation are not equivalent to criterion-referenced testing. Criterion-referenced tests (CRTs) can be used in any of the approaches we have described. They can even be used in the norm-referenced approach to impact evaluation if norms are developed for them (Roudabush, 1975).

In a criterion-referenced approach to impact evaluation, performance or status at the end of the program is compared to some clearly defined standard or criterion. Criteria may be defined in terms of some broader set of items or attributes, in what might be called a domain-referenced approach, or in terms of some directly stated goal or objective, in what might be called an objectives-based approach. In a preschool Title I program, for example, children's status might be compared with the domain-referenced criterion that each should be able to read out loud any sample of ten letters of the alphabet; or performance might be compared with the objective for the program that each should be able to count aloud from one to ten.

By comparing participants' performance at the start and the end of the program, one can derive an estimate of how much they have changed over the course of the program.* Such an estimate is, however, a very uncertain

* The criterion-referenced approach can, of course, be used in assessing performance only at the end of a program. Such end-of-program only assessment can certainly be useful for a variety of purposes, but unless a start-of-program assessment is also used (or some other basis, like a comparison group, is available for estimating change over the course of the program) this approach cannot properly be termed an impact assessment.

measure of program impact. Since this approach is just a pretest/posttest comparison, changes in criterion performance may actually be due to influences other than the program. Young children may change over the course of a program for many reasons--for example, natural maturation, instruction they receive outside the program, or any of a number of other influences. Nevertheless, some summary of children's status at the end of the program compared to their performance at the start, in terms of some clearly defined criterion, can still provide a rough indication of program performance, if not of impact.

The main advantage of this approach is flexibility. No control or comparison group is required. Any sort of outcome of interest can be encompassed, and if the program is an individualized one, different performance criteria can easily be used for different program participants.* Estimates of program impact on particular individuals can also be derived from other approaches, but with the criterion-referenced approach different outcome measures can more readily be used with different individuals.

The main weaknesses of this approach is that it is not very rigorous. Change in criterion performance between the start and end of the program may be a program effect, but may also simply reflect children's natural maturation. Unlike the approaches which attempt to adjust for the effects of maturation and other influences, the simple criterion approach provides no way to distinguish the effects of the program from changes due to other features. If the criterion performance is one which children naturally tend to improve on as they grow older--for example, in the number of words they

* This practice would, of course, raise problems of how to aggregate individual results to estimate program impact. We will discuss aggregation issues in the next chapter.

can read--then this approach will tend to overestimate program effects. The only way to make the criterion approach more rigorous, and thus more accurate in estimating program effects, is to combine it with one of the other approaches discussed. Using the criterion-referenced approach with an appropriate comparison group, for example, can help differentiate program effects from other sources of influence on children's criterion performance, or the use of growth projections as in the value-added approach can help to differentiate maturation effects from changes attributable to the program.

A second weakness of the criterion approach has to do with valid and reliable measurement. Almost any program goal or outcome of interest can be expressed in criterion-performance terms. If an ECT-I program includes social development as a goal, for example, one can develop criterion-performance measures of social development, say in terms of teachers' ratings. Yet this very flexibility may camouflage measurement problems. Are such ratings reliable and valid? This, after all, is essential with respect to any assessment. (For more information on this point, see the resource book Assessment in Early Childhood Education by Haney & Gelberg, 1980.)

Despite these problems, criterion-referenced approaches to evaluating short-term program impact can still be useful (Bryk, 1978). Program effects summarized in terms of percentages of children reaching program objectives or being "at criterion," for example, may mean more to some potential users of evaluation information than more technically sophisticated evaluation results.

In summary, like all the other approaches to estimating short-term ECT-I program impact, the criterion-referenced approach has both strengths and weaknesses. Its major strength is its flexibility. It is, for example, the most practical approach to use when different outcomes are to be assessed for different children within an ECT-I program. The major weakness of the approach is that it simply does not provide for a very strong means for telling the difference between actual program impact and other extraneous influences which may affect children's criterion performance at the end of the program. Given its flexibility the criterion-referenced approach may be more appropriate for the prekindergarten and kindergarten levels of ECT-I, since programs at these levels more often than those for older students tend to have individualized goals for different children.

VIII. AGGREGATION OF RESULTS ACROSS PROJECTS

So far, we have said little about how impact evaluation results may be aggregated across ECT-I projects. For many of the reasons cited earlier, aggregation of results with ECT-I programs raises special problems. ECT-I programs tend to have more diverse goals than later-grade Title I programs, which usually emphasize reading, mathematics, and language arts achievement. Assessment techniques at the early childhood level tend to be far more diverse than the achievement testing more commonly used with older children. And national norms--which form the basis for aggregation in later-grade Title I programs--simply are not available for many common goals of early childhood programs. For these reasons, we discuss the aggregation of results of ECT-I impact evaluations separately in this chapter.

The first question to consider in trying to aggregate or compare evaluation results across ECT-I projects is the same one that should be asked about any ECT-I impact evaluation: why do it? It may be to compare the effectiveness of different ECT-I program approaches, or to provide an accounting to various agencies or parent groups, or for some other reason. But whatever the case, the intended use and prospective users of the aggregated results should influence how one goes about aggregation of impact evaluation results.

Any aggregation effort will have to deal with three issues:

- The designs of the evaluations whose results one tries to aggregate
- The content of the outcomes across which aggregation is to be performed

- The metric that is to be used.

Design. As suggested in foregoing chapters, approaches to assessing the short-term impact of ECT-I programs vary considerably in rigor. The control group approach, if implemented properly, can give the strongest conclusions on program impact. The criterion-referenced approach generally will give the least trustworthy estimates of program impact. Thus, in planning any comparison or aggregation of results across ECT-I programs, one should keep in mind that results may differ not just because of differences in program, but because of differences in evaluation design and implementation. This, of course, applies to any effort to aggregate the results of evaluations of different design, but there is some empirical evidence that the problem may be more severe for ECT-I programs than for later-grade Title I programs* because of the special difficulty of assessing the impact of early childhood programs, as discussed above.

Content. A second key issue concerns the content of the outcomes across which one tries to aggregate results. It is commonly assumed that basic types of standardized tests (reading, mathematics, and language arts) at particular grade levels cover essentially the same content. That assumption is increasingly being challenged with respect to later-grade tests (e.g., Porter, et al., 1978); but it is often especially questionable

* Loveridge and Carapella (1979) compared the results of applications of USOE evaluation models A, B, and C to data on kindergarten Title I projects in St. Louis, Missouri. They found that effect estimates from the different models, even though based on the same data, varied by as much as 10 to 17 NCEs. A similar study by Faddis, Arter, and Zwertchek (1979), comparing results for models A and B with data from a ninth-grade Title I project, found effect estimates to differ by only 0 to 4 NCEs.

with respect to some early childhood tests which go by similar names. Rude (1973), for example, has shown that five of the more widely used reading readiness tests actually encompass very different sets of skills. Conversely, he has observed that "disagreement is also apparent in the labeling of the subtests, even though the tests are essentially similar" (p. 575). Thus, in considering whether to aggregate results across ECT-I impact evaluation studies, one must examine the content equivalence of outcomes not just in terms of the titles given to assessment instruments, but also in terms of the actual skills that particular instruments tap.

Metric. Yet another problem in trying to aggregate results across evaluations is what metric can be sensibly used. For the ED system of evaluation for later-grade Title I programs, a common reporting scale derived from test norms, namely normal curve equivalents or NCEs, has been developed. Given the special problems in norms for many early childhood tests (norms altogether missing, based on nonequivalent norm groups, or differentiated in terms of children's previous experience), this approach may not prove feasible for many early childhood outcomes. Thus, it may prove reasonable to aggregate results only across evaluation studies that use the same instrument, basing aggregation on the specific metric available for that test.

An alternative, of course, is to aggregate results in a metric-free manner. For example, with criterion-referenced approaches it may be possible to aggregate results not in terms of any independent metric, but rather on the basis of proportions of participants reaching particular criterion levels. This is exactly what some states have done in aggregating Title I evaluation results (e.g., West, 1976.) Also, it may be possible, under certain assumptions, to transform other sorts of evaluation results into criterion-referenced form--for example, proportion of children scoring above the twentieth percentile at posttest time.

In sum, short-term impact.evaluation results generally will prove more difficult to aggregate across different ECT-I programs than across later-grade Title I programs. Any such aggregation must be planned in light of the purposes and persons one hopes to inform, and with special attention to the design, the content of measures, and the metric employed in the individual short-term impact evaluations across which one wishes to aggregate.

IX. SUMMARY AND CONCLUSIONS

In Chapters III-VII of this booklet, we reviewed five approaches to estimating short-term program impact:

- Norm-referenced approach
- Comparison group approach
- Regression approach
- Value-added approach
- Criterion-referenced approach.

These approaches to estimating program impact differ both in their general characteristics and in the potential problems they raise. This concluding chapter

- Summarizes the strengths and weaknesses of these approaches from these two perspectives
- Briefly recaps the technical standards mandated by ED for all Title I evaluations
- Recounts the issues which must be addressed in any effort to aggregate impact evaluation results across programs
- Suggests some alternative ways to think about the purposes served by these approaches to Title-I or any other evaluation.

General issues. The five approaches to estimating program impact can in principle be applied to evaluate the effectiveness of any program, educational or otherwise. The approaches differ substantially, however, in two ways: (1) their practical requirements; and (2) the quality or validity of the inferences they can yield with respect to impact or effect estimation. Some of these general characteristics are summarized in Figure 4. As the figure suggests, some important trade-offs are implicit in the different approaches. As a general rule, the easier an approach is to implement--that is, the fewer practical requirements it has--the lower will be the quality or validity of the inferences or conclusions which can be drawn from it. Conversely, approaches that yield more clear-cut conclusions generally carry with them more constraints in terms of practical requirements. By and large, the comparison-group, regression, and value-added approaches will be more difficult to implement properly, but will yield relatively stronger conclusions or inferences about program impact. The norm-referenced and criterion-referenced approaches generally will be easier to implement but will yield weaker or lower-quality inferences regarding program impact.

This pattern is, however, only a very loose one. In practice, the quality of inferences to be drawn will depend mainly on how each approach is applied. The criterion-referenced approach, for example, may in general fail to control for such influences as children's natural growth, or their experiences outside the program, but if thoughtfully applied may nevertheless yield more valid conclusions than, say, a comparison group evaluation that is badly done.

As Figure 4 suggests, none of these approaches is likely to result in estimates of program impact in which one should have very strong confidence. Instead it is in general more appropriate to view the results of any one evaluation of an ECT-I program's short-term impact as merely suggestive. Results of short-term impact evaluation may be more valuable if combined with other evaluation strategies as suggested below. The point that technical issues do not determine an evaluation's worth can be illustrated by describing briefly an evaluation which surpassed by far the technical sophistication of any local Title I evaluation. We refer to the national evaluation of Project Follow Through. FT, like Title I, is a compensatory education

Approach	Explanatory Value	Comparison or Control Group Required	Norm-Referenced Assessment Required	Complex Statistical Calculations Required	Fairly Large Sample of Participants Required (30 or more)	Quality of Inferences Drawn
Norm-Referenced	Weak	No	Yes	No	No	Weak
Comparison Group	Moderate	Yes	No	Yes*	Yes	Moderate
Regression	Strong	Yes	No	Yes	Yes	Weak to Moderate
Value-Added	Strong	No	No	Yes	Yes	Weak to Moderate
Criterion-Referenced	Moderate	No	No	No	No	Weak

*As normally applied, the comparison group approach requires statistical calculations to adjust for differences between control group and program participants. However, if control and program participants are selected in highly similar ways, such statistical adjustments may be unnecessary.

Figure 4. Summary of Characteristics of Five Approaches to Estimation of ECT-I Program Impact.

program aimed at improving the learning of educationally disadvantaged students. The national evaluation of FT was a massive effort, lasting more than ten years and costing around $50 million. In many ways the impact evaluation of FT was far more technically sophisticated than previous impact evaluations of education programs (Haney, 1977). Several different kinds of comparison groups and numerous complex statistical analyses were employed in estimating program effects. Nevertheless, the FT evaluation results became embroiled in considérable controversy, were publicly challenged as being technically deficient (House et al., 1978) and apart from providing grist for debate among evaluation specialists had little value in terms of program improvement. The example clearly indicates that technical sophistication is simply not enough to guarantee the utility of our evaluation.

Special issues with respect to early childhood. Whatever their virtues from a technical point of view, each of the five approaches to short-term impact evaluation may raise special issues when applied with respect to ECT-I programs.

The main strength of the norm-referenced and criterion-referenced approaches is that they can be fairly simple to implement. Also, the norm-referenced approach may be attractive for some school-based ECT-I programs (at the kindergarten and grade-1 levels), for the simple reason that norm-referenced tests may already be regularly administered in school testing programs. On the other hand, norm-referenced tests may not be available for some goals of ECT-I programs, and norm results may be affected sharply by children's previous educational experiences. Indeed, the major general weaknesses of both these approaches is that they provide little means for differentiating program impact from extraneous factors affecting children's growth and performance.

The three other approaches--comparison group, regression and value-added--provide somewhat broader bases for differentiating actual program impact from other influences, but the means by which each controls for such influences also can be a source of problems. In the comparison group approach, the control group provides the basis for estimating the no-treatment expectation, hence controlling for extraneous influences. But for such control to be effective, it is crucial that the comparison group be similar to the program participants in all respects other than that they do not receive ECT-I services. If such a comparison group is not available then this approach cannot be applied to ECT-I programs. With the regression approach a control or comparison group also provides the basis for estimating the no-treatment expectation, but in this case via statistical computations concerning the relationship between pretest and posttest scores. Problems that can arise in such calculations with respect to ECT-I programs may derive from the nature of the test or other assessment instrument used. If a test is too hard or too easy for children (what are often called floor and ceiling effects, respectively) at either pretest or posttest time, then the statistical calculations may not work out properly.

Similar complications may arise with the value-added approach. This approach capitalizes on the relationship between children's ages and pretest scores to estimate no-treatment expectation without resorting to use of a control or comparison group. But if no such relationship exists--whether becuase of the nature of the attribute of interest, measurement problems like floor or ceiling effects, or they way program participants were selected--then this approach may not be possible to apply.

Technical standards. Having reviewed five different approaches to estimating short-term impact of ECT-I programs, let us briefly summarize four general technical standards which ED has mandated with respect to all Title I evaluation efforts. First, an impact evaluation should employ valid assessment of program goals. Second, evaluations should be implemented so as to assure that findings are representative of the whole program; that is, they are based on all or a representative sample of individuals served in the program. Third, evaluation instruments and procedures should be both reliable and valid. Fourth, quality control procedures should be instituted so as to minimize errors in data gathering, analysis and reporting.

Aggregation. Under some circumstances, one may wish to aggregate short-term impact evaluation results across more than one ECT-I program or across more than one program period. There are three basic issues which must be considered in doing so. First, are the designs of the evaluations comparable? If not, that is, if different evaluations control for different possible influences extraneous to program impact, then different results may simply represent different qualities of the designs employed. Second, the different outcomes across which one wishes to aggregate results must represent the same or highly similar content. Third, one must consider whether a common metric is available or can be developed. Since for many important outcomes of ECT-I programs no appropriate norm-referenced tests are available, norm-referenced scales (such as NCEs, stanines, or national percentiles) may not be possible to use. In such cases it may be necessary to use raw test scores or some empirical or content equating of the outcome measures. If no such metric is available or can reasonably be developed, then the only alternative is to employ a metric-free method of aggregating results. For example, if different outcomes can be expressed in criterion terms, say in terms of percentages of children reading at specified levels of attainment, then it may be possible to compare results across different evaluations in these terms.

Whether to attempt aggregation of results across different ECT-I evaluations should not, of course, depend exclusively on technical considerations. Instead, the key question to be asked here, as with most other evaluation issues, is what will make sense in light of the evaluation purposes one is aiming to serve. Ultimately, the test of the value of short-term impact evaluation of an ECT-I program rests not on technical issues, but rather on how well the evaluation contributes to better decision making and improved ECT-I programs.

Purposes of evaluations. In conclusion, it is important to repeat that short-term impact evaluations of ECT-I programs should not be viewed in isolation or as simply a formal reporting requirement. One must consider not just whether an impact estimate was derived or whether an evaluation report was produced but also whether the information derived from an impact evaluation yielded a better understanding of how the program works and how it can be improved in the future. In this light it is important to consider impact evaluation not simply as a technical undertaking, but more broadly as one among many means of learning how ECT-I programs operate. As such, impact evaluations often may prove most informative if combined with other methods of evaluation (see Apling and Bryk, 1980, for a discussion of how impact on outcome evaluations can be combined with other evaluation strategies).

NOTES ON SOURCES OF FURTHER INFORMATION

General Sources. A good, easily available introduction to the ED models for evaluating the impact of Title I programs in grades 2-12 is Tallmadge & Wood (1978, 1980). For background information on the history and design of the ED Title I evaluation system, see Wisler & Anderson (1979) and Cross (1979). For critical comments on the Title I evaluation system, see Linn (1979), Jaeger (1979), and Wiley (1979). The last five references come from a single issue of the journal Educational Evaluation and Policy Analysis (1:2, March-April 1979). A good source of information on more general issues in the evaluation of early childhood education programs is Goodwin and Driscoll's (1980) Handbook for Measurement and Evaluation in Early Childhood Education.

Regional Title I Technical Assistance Centers (TACs) are an excellent source of information on a range of practical and technical issues concerning Title I evaluation (see list at back of this booklet).

Norm-Referenced Approach. A general description of the norm-referenced model of impact evaluation for later-grade Title I programs is Tallmadge & Wood (1980), Chapter 4. On the use of local norms, see Wood & Tallmadge (1976).

One of the most commonly discussed technical problems with respect to the norm-referenced model is the statistical regression effect which tends to cause students' test scores to increase or decrease upon retesting simply because of measurement error and the way students are selected. For example, if 100 students are tested, and the lowest-scoring 25 selected for retesting, they can be expected as a group to score higher on the retesting simply because of measurement error. To help overcome the regression effect in application of the norm-referenced model, ED has recommended that in applications of the norm-referenced approach either participants in Title I projects not be selected on the basis of pretest scores or that a statistical correction formula be applied to adjust for the regression effect. Analyses by Echternacht (1979) indicate that use of different tests for selection and pretesting does not, in general, eliminate the regression effect. Data cited by Echternacht indicate that the regression effect may be more pronounced at lower grade levels, apparently because tests of younger children tend to carry larger degrees of measurement error than those of older children.

Analyses of standardized growth expectancies by Stenner et al. (1978) illustrate both the rapid rate of development of young children in terms of norm-referenced test results and how norm-referenced results may be very misleading when comparisons are drawn between the early childhood and later grade levels.

More general information on test norms can be found in Anastasi (1976), and technical information concerning test norms is available in Angoff (1971).

Comparison Group Approach. The comparison group model for short-term impact evaluation, as it applies to later-grade Title I evaluation, is described in Tallmadge & Wood (1980), Chapter 5. An introduction to use of statistical adjustments with respect to nonequivalent control groups can be found in Tallmadge & Horst (1976). For more thorough technical treatments of the same topic, see Kenney (1975) and Bryk & Weisberg (1977).

Regression Approach. .Tallmadge & Wood (1980), Chapter 6, discusses implementation of the regression model as it applies to impact evaluation of later-grade Title I programs. This document recommends estimation of program effect as observed minus expected treatment group posttest means, since effects estimates may vary in terms of the point at which they are estimated when treatment and comparison group regression lines are not strictly parallel.

For a fuller discussion of the regression approach to estimation of Title I program impact in general, see Echternacht and Swinton (1979). They conclude:

1. Although model C works well when there are no floor or ceiling effects present, nonlinearities may have a large effect on the calculation of impact estimates.

2. In addition to the usual model C analysis, evaluators should apply a parallel-slopes fit and any other fits that seem reasonable, and compare results. If parallel slope and model C procedures give similar estimates, curvilinearity is probably not serious.

3. In any case the interocular impact test (does it hit you between the eyes?) is always advisable. This test requires graphing the scatter plot of pre vs posttest scores and LOOKING.

Value-Added Approach. A good general introduction to the value-added approach to the measurement of short-term ECT-I program impact is Bryk and Woods (1980). More technical information on this approach to estimating program impact can be found in Bryk & Weisberg (1974, 1976, and 1977) and Strenio (1977). Examples of the application of this approach in estimating the short-term impact of Head Start programs may be found in Smith (1973) and Weisberg (1974). For an example of an application of the value-added approach to estimating program effects with older children see Messick (1980).

Criterion-Referenced Approach. Criterion-referenced approaches to short-term impact evaluation are based essentially on three steps. First, specific objectives are defined in terms of skills or behaviors which a program seeks to impart to program participants. Second, some means of assessing whether participants have reached objectives is selected or developed and tried out (and refined as necessary). Third, participants are assessed at the end of the program and program "success" is summarized as the proportion of participants reaching objectives or "at criterion." As mentioned in the text, such a summary is only a crude indicator of program impact unless this approach is combined with one of the other approaches to estimating impact. For a discussion of the general strengths and weaknesses of such a goal- or objectives-based approach to evaluation, see Popham (1974), especially pp. 34-67. For a discussion of objectives-based evaluation with respect to early childhood education, see Goodwin & Driscoll (1980), pp. 346-349.

Aggregation. For critiques of the potential aggregation of results across the three later-grade Title I evaluation models, see Wiley (1979) and Echternacht (1978). Two good sources on the general topic of aggregation of results across different impact evaluation studies are Pillemer and Light (1980) and Glass (1977). For more general information on the topic of aggregation in data analysis, see Roberts & Burstein (1980).

Title I Technical Assistance Centers. The Title I Technical Assistance Centers serving the ten regional areas of the United States are good sources of up-to-date information on Title I evaluation.

Region I: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont

-RMC Research Corporation
400 Lafayette Road
Hampton, N.H. 03842
Telephone: (603) 436-5385
926-8888

<u>Region II</u>: New York, New Jersey, Puerto Rico, and the Virgin
Virgin Islands

-Educational Testing Service
Princeton, N.J. 08540
Telephone: (609) 734-5117

<u>Region III</u>: Delaware, Maryland, Pennsylvania, Virginia,
West Virginia, and the District of Columbia

-National Testing Service
2634 Chapel Hill Blvd.
Durham, N.C. 27707
Telephone: (919) 493-3451
(800) 334-0077

<u>Region IV</u>: Alabama, Florida, Georgia, Kentucky, Mississippi,
North Carolina, South Carolina, and Tennessee

-Educational Testing Service
Southern Regional Office
250 Piedmont Avenue
Suite 2020
Atlanta, Georgia 30326
Telephone: (404) 524-4501

<u>Region V</u>: Illinois, Indiana, Michigan, Minnesota,
Ohio, and Wisconsin

-Educational Testing Service
1 American Plaza
Evanston, Illinois 60201
Telephone: (312) 869-7700

<u>Region VI</u>: Arkansas, Louisiana, New Mexico, Oklahoma,
and Texas

-Powell Associates
3724 Jefferson
Suite 205
Austin, Texas 78731
Telephone: (512) 453-7288
(800) 531-5239

<u>Region VII</u>: Iowa, Kansas, Missouri, and Nebraska

-American Institutes for Research
P.O. Box 1113
Palo Alto, CA 94302
Telephone: (415) 494-0224

<u>Regions VIII, IX and X:</u> Colorado, Montana, North Dakota, South Dakota, Utah, and Wyoming (Region VIII); Arizona, California, Hawaii, Nevada, Guam, Trust Territory of the Pacific Islands, and American Samoa (Region IX); and Alaska, Odaho, Oregon, and Washington (Region X)

Northwest Regional Laboratory
300 S.W. Sixth Avenue
Portland, Oregon 97204
Telephone: (503) 295-0214

REFERENCES

Anastasi, A. Psychological Testing (Fourth Edition). New York: MacMillan, 1976.

Anderson, J.K, Johnson, R.T., Fishbein, R.L., Stonehill, R.M., & Burnes, J.C. The U.S. Office of Education Models to Evaluate ESEA Title I: Experiences After One Year of Use. Washington, D.C.: USOE, Office of Planning, Budget, and Education, 1978.

Angoff, W. Scales, norms and equivalent scores. In R. Thorndike (ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1978.

Apling, R., & Bryk, S. Evaluation Approaches: A Focus on Improving Early Childhood Title I Programs. Cambridge, MA: The Huron Institute. 1980

Bryk, A.S. Evaluating program impact: A time to cast away stones, a time to gather stones together. In S. Anderson (ed.), New Directions for Program Evaluation. San Francisco, CA: Jossey-Bass, 1978.

Bryk, A., Apling, R., & Mathews, R. Developing an Evaluation System for Early Childhood ESEA Title I: A Feasibility Analysis (Draft). Cambridge, MA: The Huron Institute, 1978.

Bryk, A., & Weisberg, H. A new approach to analyzing quasi-experimental data. Proceedings of the Social Statistics Section, American Statistical Association, August 1974.

Bryk, A., & Weisberg, H. Value-added analysis: A dynamic approach to the estimation of treatment effects. Journal of Educational Statistics, 1976, 1:2, 127-155.

Bryk, A., & Weisberg, H. Use of the nonequivalent control group design when subjects are growing. Psychological Bulletin, 1977, 84:5, 950-962.

Bryk, S., & Woods, E. An Introduction to the Value-Added Model and Its Use in Short-Term Impact Assessment. Cambridge, MA: The Huron Institute, 1980.

Cronbach, L., & Snow, R. Aptitudes and Instructional Methods. New York: Irvington, 1977.

Cross, C. Title I evaluation: A case study in congressional frustration. Educational Evaluation and Policy Analysis, 1979, 1:2, 15-21.

Darlington, R. Preschool programs and later school competence of children from low income families. Science, 1980, 208, 202-204.

Echternacht, G. The Norm Referenced Model and the Regression Effect. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA., April 8-12, 1979.

Echternacht, G. The Use of Different Models in the ESEA Title I Evaluation System. Unpublished paper, Princeton, N.J.: Educational Testing Service, March 1978.

Echternacht, G. & Swinton, S. Getting Straight: Everything You Always Wanted to Know About the Title I Regression Model and Curvilinearity. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA., April 8-12, 1979.

Faddis, B., Arter, J., & Zwertchek, A. An Empirical Comparison of ESEA Title I Evaluation Models A and B. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.; April 8-12, 1979.

Financial assistance to local educational agencies to meet the special educational needs of educationally deprived, and neglected and delinquent children--evaluation requirements. Proposed rules. Federal Register, 1979, 44:27 (February 7, 1979), 7914-7919.

Financial assistance to local educational agencies to meet the special educational needs of educationally deprived and neglected and delinquent children--evaluation requirements. Final regulations. Federal Register, 1979, 44:199 (October 12, 1979), 59152-59159.

Glass, G. Integrating findings: The meta-analysis of research. Review of Educational Research on Education, 1977, 5, 351-379.

Goodwin, W. and Driscoll. L. Handbook on Measurement and Evaluation in Early Childhood Education. San Francisco, CA.: Jossey-Bass, 1980.

Haney, W. The Follow Through Planned Variation Experiment. Vol V: A Technical History of the National Follow Through Evaluation. Cambridge, MA: The Huron Institute, 1977.

Haney, W., & Gelberg, W. Assessment in Early Childhood Education. Cambridge, MA: The Huron Institute, 1980.

House, E., et al. No simple answer: A critique of the Follow Through evaluation. Harvard Educational Review, May, 1978, 48:2.

Jaeger, R. The effect of test selection on Title I project impact. Educational Evaluation and Policy Analysis, 1979, 1:2, 33-40.

Kennedy, M. Longitudinal Information Systems in Early Childhood Title I Programs. Cambridge, MA.: The Huron Institute, 1980.

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

Kodlin, P., & Thompson, D.J. An appraisal of the longitudinal approach to studies of growth and development. Monographs of the Society for Research in Child Development, 1958, 23(1), Series No. 67.

Lazar, I., Hubbell, V., Murray, H., Rosche, M., & Royce, J. The Persistence of Preschool Effects (Report for the Education Commission of the States). Ithaca, NY: Community Service Laboratory, NYS College of Human Ecology, Cornell University, 1977.

Linn, R. Validity of inferences based on the proposed Title I evaluation models. Educational Evaluation and Policy Analysis, 1979, 1:2, 23-32.

Loveridge, R., & Carapelle, R. The Application of USOE Models A, B, and C to a Title I Early Childhood Program. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, April 8-12, 1979.

Messick, S. The Effectiveness of Coaching for the SAT: Review and Reanalysis of Research from the Fifties to the FTC. Princeton, N.J.: Educational Testing Service, 1980.

Pedersen, E., & Faucher, T. A new perspective on the effects of first-grade teachers on children's subsequent adult status. Harvard Educational Review, 1978, 48, 1-31.

Pillemer, D., & Light, R. Synthesizing outcomes: How to use research evidence from many studies. Harvard Educational Review, 1980, 50:2, 176-195.

Popham, J. (ed.). Evaluation in Education. Berkeley, CA.: McCutchan, 1974.

Porter, A., et al. Practical significance in program evaluation. American Research Journal, 1978, 15:4.

RMC Research Corporation. Policy Manual Chapter 10 Evaluation (draft). Mountain View, CA: Author, September 1980.

Roberts, K., & Burstein, L. (eds.) Issues in Aggregation. San Francisco, CA: Jossey-Bass, 1980.

Roudabush, G. Estimating Normative Scores from a Criterion-Referenced Test. Paper presented at the Annual Meeting of the Educational Research Association, Washington, D.C., April 1975.

Rude, R.T. Readiness tests: Implications for early childhood education. The Reading Teacher, 1973, 26(6), 572-580.

Scriven, M. Evaluation perspectives and procedures. In J. Popham (ed.). Evaluation in Education. Berkeley, CA: McCutchan, 1974.

Smith, M.S. Some Short-Term Effects of Project Head Start: A Preliminary Report on the Second Year of Planned Variation: 1970-71. Cambridge, MA: The Huron Institute, 1973.

Stenner, A.J., Hunter, E.L., Bland, J.D. and Cooper, M.L. The Standardized Growth Expectation: Implications for Educational Evaluation. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, 1978.

Strenio, J.F. An individual growth model perspective for evaluating educational programs (Qualifying Paper, Harvard Graduate School of Arts and Sciences), 1977.

Tallmadge, K., & Horst, D. Statistical Adjustments for Nonequivalent Control Groups. ESEA Title I Evaluation and Reporting System Technical Paper No. 12 (Draft). Mountain View, CA: RMC Research Corporation, 1976.

Tallmadge, K., & Wood, C. User's Guide: ESEA Title I Evaluation and Reporting System (Revised). Mountain View, CA: RMC Research Corporation, 1978 [Draft, revised, September 1980].

Weisberg, H. Short-Term Cognitive Effects of Head Start Programs: A Report on the Third Year of Planned Variation: 1971-72. Cambridge, MA: The Huron Institute, 1974. (ERIC No. ED 093 047.)

Weisberg, H., & Haney, W. Longitudinal Evaluation of Head Start Planned Variation and Follow Through. Cambridge, MA: The Huron Institute, 1977.

West, R. FY '76 Evaluation Report on Title I 89-10. Illinois: State Board of Education, Illinois Office of Education, November 1976.

Wiley, D. Evaluation by aggregation: Social and methodological biases. Educational Evaluation and Policy Analysis, 1979, 1:2, 41-45.

Wisler, C., & Anderson, J. Designing a Title I evaluation system to meet legislative requirements. Educational Evaluation and Policy Analysis, 1979, 1:2, 47-55.

Wood, C., & Tallmadge, K. Local Norms. ESEA Title I Evaluation and Reporting System Technical Paper No. 7 (Draft). Mountain View, CA: RMC Research Corporation, 1976.

Yurchak, M.J., & Bryk, A.S. ESEA Title I Early Childhood Education: A Descriptive Report. Cambridge, MA: The Huron Institute, 1978.

*U S GOVERNMENT PRINTING OFFICE: 1981 728-189/1791