DOCUMENT RESUME

ED 199 340 UD 021 279

AUTHOR DiCostanzo, James L.; Eichelberger, R. Tony
TITLE Reporting Results in Evaluation Settings: Emphasizing Selected Issues in ANCOVA Analysis and Interpretation.
INSTITUTION Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO LRDC-1979/14
PUB DATE 79
NOTE 47p.; Not available in paper copy due to reproduction quality of original document.

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Analysis of Covariance; *Data Collection; Elementary Secondary Education; *Evaluation Criteria; *Evaluation Methods; Information Needs; *Program Evaluation; *Research Methodology; Research Problems
IDENTIFIERS *Project Follow Through

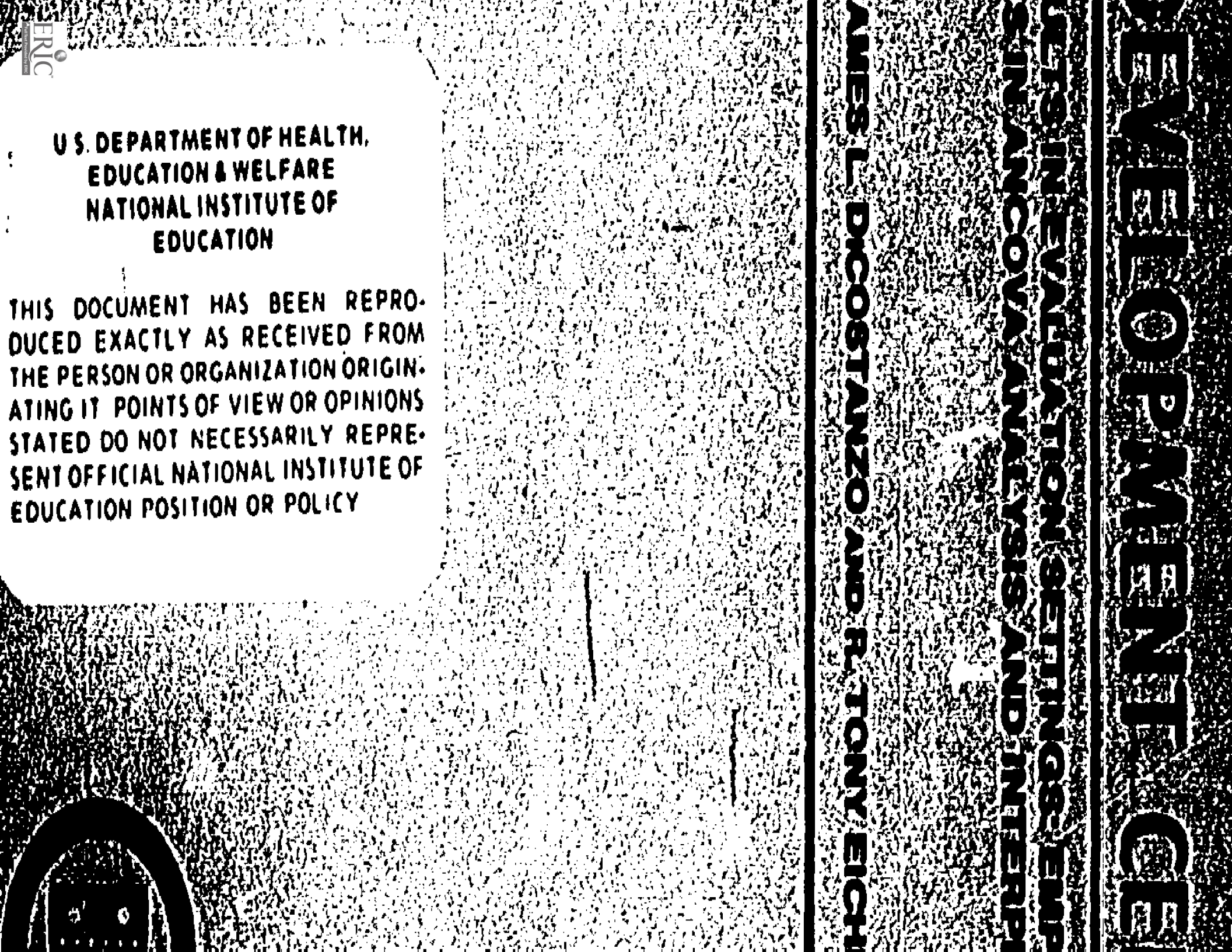
ABSTRACT

Evaluators often utilize analysis of covariance (ANCOVA) techniques to compensate statistically for the lack of experimental control when assessing the effects of innovative programs implemented in naturalistic settings. In this paper design, analysis, and reporting considerations important to the application of ANCOVA-type techniques in educational settings are described. Problems that arise from the use of complex data analysis techniques are identified, based on a review and critique of the evaluation of the national Follow Through program. Specific information that should be included in an evaluation report when ANCOVA-type techniques are used is described. Examples of the kinds of problems that appear when collecting data in school settings are provided in order to illustrate the need for this information. Alternative ways of presenting the needed information in an evaluation report are discussed. The overall perspective is that evaluation reports must be more precise and indicate the limitations as well as strengths of the methodology used for the specific setting. (Author/APM)

LEARNING RESEARCH

REPO
SELE

1979/14



REPORTING RESULTS IN EVALUATION SETTINGS: EMPHASIZING SELECTED ISSUES IN ANCOVA ANALYSIS AND INTERPRETATION

James L. DiCostanzo and R. Tony Eichelberger

Learning Research and Development Center

University of Pittsburgh

1979

The preparation of this paper was supported by the Learning Research and Development Center, supported in part as a research and development center by funds from the National Institute of Education (NIE), United States Department of Health, Education, and Welfare. The opinions expressed do not necessarily reflect the position or policy of NIE, and no official endorsement should be inferred.

Abstract

Evaluators often utilize ANCOVA-type techniques to assess the effects of innovative programs implemented in naturalistic settings. In this paper design, analysis, and reporting considerations important to the application of ANCOVA-type techniques in educational settings are described. Numerous examples are drawn from the national Follow Through evaluation, and suggestions for improving reports utilizing such ANCOVA-type techniques are presented. The overall perspective is that evaluation reports must be more precise and must indicate the limitations as well as the strengths of the methodology used for this specific setting. In doing so, a more balanced description of a program and its effects is presented to the decision maker and to other stakeholders.

REPORTING RESULTS IN EVALUATION SETTINGS: EMPHASIZING SELECTED ISSUES IN ANCOVA ANALYSIS AND INTERPRETATION

James L. DiCostanzo and R. Tony Eichelberger

Learning Research and Development Center
University of Pittsburgh

Since the early 1960's, the federal government has authorized and funded numerous social action programs, many of which focused on compensatory education. The evaluations of these programs have usually been attempts to implement an experimental paradigm designed to maximize internal validity. Since manipulation of important variables is rarely possible (and often not appropriate) in evaluation settings (Cooley, 1978), some type of analysis of covariance (ANCOVA) technique is frequently utilized to compensate statistically for the lack of experimental control.

Use and interpretation of the ANCOVA technique is extremely complex, requiring that numerous assumptions and conditions be met if meaningful interpretations are to be applied to educational settings. These assumptions are never precisely met in an evaluation setting, so the extent of the deviations and their impact on meaningful interpretations must be assessed and presented in the evaluation.

The types of problems that arise from the use of complex data analysis techniques, such as ANCOVA, that are addressed in this paper were identified from a review and critique of the evaluation of the national Follow Through program. There has been no attempt to comprehensively identify the evaluation problems or the issues that relate to utilizing and reporting ANCOVA results. The problems and issues discussed are of recurrent concern to educational evaluators in various settings.

In this paper, specific information that should be included in an evaluation report when ANCOVA-type techniques are used is identified. This information should enable the reader to accurately assess the adequacy of the technique and the appropriateness of the evaluator's interpretation of the results for that particular setting. Specific examples of the kinds of problems that arise when collecting data in school settings are described to illustrate the need for this additional information. Alternative ways of presenting the needed information in an evaluation report are presented and discussed.

The comments and suggestions made in this paper follow primarily from the longitudinal evaluation of the national Follow Through program. This evaluation is a typical example of the application of an ANCOVA-type analysis technique in an evaluation setting. Because of its scope and duration (six years), the Follow Through evaluation encountered most of the problems that evaluators must face and that result from the use of this technique.

National Follow Through Program

A brief historical sketch of the national Follow Through program and its changing purposes is needed to understand and appreciate the methodological issues discussed in the remainder of this paper. In 1966, there were indications that Head Start, a federally funded compensatory education program for disadvantaged preschool children, was having some positive effects, but that the effects did not endure through the early elementary school years (Wolff & Stein, 1966). The Follow Through (FT) program was planned as a massive service program and was designed to extend compensatory education (similar to that afforded the Head Start children) from kindergarten through grade three (Johnson, 1967). When FT was originally funded, only $15 million was appropriated for two years, rather than the $120 million that was expected.

To the Office of Education, Follow Through then became a planned variation experiment in which diverse types of innovative programs were implemented in various sites throughout the U. S. But, rather than assigning programs randomly to sites or projects as in a controlled experiment, participating local districts, in cooperation with the programs' sponsors, were allowed to select the instructional model to be implemented in their project. Although this procedure later caused some methodological problems, it is probably more representative of the operation of U. S. public schools than is the random assignment of programs to sites.

In the initial two years of the FT program (1967-68 and 1968-69), the evaluation focus was somewhat confused, due primarily to the change in the program emphasis from service to a planned variation experiment and the associated administrative problems. In 1968-69, several purposes for the national Follow Through evaluation were delineated, including: (a) assessing program impact on pupils, parents, schools, and community (Emrick, Sorensen, & Stearns, 1973, p. 72); (b) assessing relative effectiveness of different programs and program approaches (Sorensen & Madow, 1969, p. 4); and (c) establishing criteria for effectiveness and success of the national FT program (Sorensen & Madow, 1969, p. 4).

In this paper, we are concerned with selected aspects of these three purposes, which deal with the impact of the FT programs. The evaluation attempted to accomplish these purposes by using ANCOVA.

Approximately 70 of the 170 local projects representing 14 of 22 FT sponsor models were included in the national FT evaluation. In each FT school district, students identified as similar to those participating in FT comprised the Non-Follow Through (NFT) sample and were tested on a regular basis by Stanford Research Institute (SRI), the organization contracted to collect all FT evaluation data. When

comparable students could not be identified locally, a comparison or control group from a neighboring school district was identified and tested. Noncomparability of the FT and NFT groups at a particular site was often a result of the school district's policy of assigning the most disadvantaged children to the FT program. Noncomparability, for this and other reasons, was an ongoing problem in the evaluation that the use of ANCOVA attempted to alleviate, despite the lack of randomization in the design.

Decision makers associated with the early years of FT were confident that the program would have a marked impact on the participating children. Richard Egbert (1973), the original FT Director, indicated that the evaluation design was based on the conviction that:

> children's development would be so markedly superior as to be readily demonstrated on measures of achievement, cognition, self-concept, social maturation, and capacity to function independently. Follow Through's design was born also from the conviction that unless such substantial differences were manifest, the really massive increases in spending that would be required could not be justified. (p. 25)

These convictions seem to have resulted in less concern with details of the design, since it was believed that any reasonable evaluation of FT would readily show the impact and effectiveness of the program.

The FT evaluation has vacillated in emphasis from a decision orientation of identifying the "best" model(s) overall to a descriptive orientation in which different effects of individual models would be described. Initially, SRI was awarded an evaluation contract to identify the most effective program model(s) and to provide descriptive information to project administrators and other school administrators. At various times, it was decided that a consumer's guide, which would list individual sponsors' objectives and the degree to which the objectives were met, was to be produced by SRI. Since 1972, the major

objective of the national FT evaluation has been to identify the successful model(s) and to document the impact of the models on pupils. An ANCOVA-type procedure has been utilized for this purpose.

SRI and Abt Associates, the major contractors for the longitudinal evaluation of the impact of FT, have produced four reports. The SRI report covered the interim years of FT, 1969-71. Abt Associates have produced four reports covering the years 1972 through 1975. The SRI report (Emrick et al., 1973) and the third Abt report (Stebbins, 1976) are used for illustrative purposes in this paper. For an extensive review and critique of the FT evaluation, see Haney (1977).

Analysis of Covariance (ANCOVA)

As indicated above, ANCOVA is often used in evaluation settings where it is difficult or impossible to control experimentally alternative explanations of educational outcomes. In situations where its use is appropriate, it allows groups to be compared on a criterion variable that has been adjusted on a set of concomitant variables, or covariates. Statistically, ANCOVA is used to increase the precision of the analysis by taking advantage of the linear relationships between the dependent variable(s) and the covariate(s). In order for ANCOVA to be unambiguously used, however, its assumptions and conditions must be precisely met. Failure to do so may distort the results in ways that make their interpretation equivocal, if not meaningless.[1]

---

[1] Abt Associates' evaluation report (Stebbins, 1976) discussed several problems associated with the analysis of data collected in the FT evaluation setting. We have selectively drawn examples from that report to illustrate our points. As a result, our paper tends to emphasize only the most questionable analysis and reporting procedures in the Abt report. Abt had the very difficult task of attempting to draw conclusions from a complex non-experimental setting. See Appendix A for a brief statement of Abt's view of their role and situation (Stebbins, 1976, p. A-46).

We believe that the consumer of the evaluation report must be able to: (a) assess the appropriateness of ANCOVA whenever it is used, and, (b) examine possible alternative interpretations of the results. For these purposes, information regarding the conformity or nonconformity to the assumptions and conditions of ANCOVA, and other information that would enable alternative interpretations of the results to be made, must be available in the report.

Areas of Concern

We have delineated some information we believe is necessary for the reader to achieve the two purposes stated above, and we have organized it into five topical areas. Each area is focused by one or more questions that the evaluator should address.

How are the Specific Research Hypotheses Investigated and the Results of the Corresponding ANCOVA Data Analyses Related to the General Evaluation Question(s)?

It is generally accepted that no empirical process completely assesses an event, and evaluation is no exception. With limited resources, especially of time, money, and personnel, an evaluation can only address some aspects of a general evaluation question.

An evaluation is defined by the specific research questions or hypotheses that are investigated. The selection of hypotheses to be tested or questions to be addressed is the result of a reasoning process that links the research hypotheses to the general question.[2]

---

[2] In an evaluation, the variables utilized are usually specified at three different levels. First, the general area of focus, such as program impact on participants or program effectiveness, is delineated. (cont.)

The explication of this reasoning process, or rationale, permits the reader of the evaluation report to identify and assess the components that are included as well as those that are not, in order to answer the general evaluation question. This explication is crucial, especially in large-scale evaluations where the inferential process relating the overall question to the specific research hypotheses is extremely complex and not obvious, especially to the reader.

One of the general impact questions specified by SRI (Emrick et al., 1973) for the national FT evaluation was, "How effective is Follow Through as a method of improving the life chances of participating children?" (p. 72). Three research questions concerned with the academic performance of FT pupils and attitudinal changes of their parents and teachers were delineated to address this general question.

How these academic performance and attitudinal change variables relate to improved life chances is not immediately apparent. A rationale that relates them is needed to enable the reader to gain an appropriate perspective for viewing the evaluation results. Cohen and Garet (1975) describe one line of reasoning in their article on social policy research:

> In the late 1950's and earl · 1960's, for example, a national policy concerning educational opportunity began to take shape. It rested partly on the idea that poverty, unemployment and delinquency resulted from the absence of particular skills and attitudes--reading ability, motivation to achieve

---

Next, the specific aspects of the area of interest that are to be investigated are specified as the questions to be addressed. Finally, each question is addressed by one or more statistical analyses. We are calling these levels: (a) the general, or overall, evaluation questions, (b) the research questions, and (c) the statistical hypotheses, which are operationalized by the actual data analyses carried out.

> in school and the like. There was also an assumption that schools inculcated these skills and attitudes and that acquiring them would lead to economic and occupational success. In other words, this policy assumed that doing well in schools led to doing well in life. (p. 21)

By specifying the rationale used, the evaluator clarifies the viewpoint on which the evaluation is based and enables the reader to understand the intentions of the evaluation. Whether or not the reader agrees with the evaluator's logic is not the important issue; we believe that scrutiny of it is necessary for the reader to assess and interpret adequately the evaluation report and the conclusions drawn from the investigation.

The need to specify the link between the research hypotheses and the overall evaluation questions has been discussed. Similarly, specifying the relationship between the statistical hypotheses actually tested and the corresponding research hypotheses is needed. Often the statistical hypotheses tested are not stated in the evaluation report. In evaluation studies or analyses that are not complex, the specific hypothesis that is tested can easily be inferred from a description of the analysis performed. This is a much more difficult task when multiple dependent and concomitant variables are analyzed or numerous analyses are used to investigate each research question.

Abt Associates' national evaluation of FT (Stebbins, 1976) is a good example of a complex evaluation utilizing numerous sophisticated analyses. An example from this evaluation that illustrates the problem and indicates an approach for dealing with it follows. One of the general evaluation questions addressed in their report was, "Does Follow Through have a greater impact on disadvantaged children than do regular school programs?" (Stebbins, 1976, p. A-8). The impact question was addressed by a number of ANCOVA analyses comparing

FT with local, best-match, and national FT groups. The results are reported in what was called Summary of Effects tables (see Table 1).

Table 1
Sample Summary of Effects Table
(Stebbins, 1976, p. A-6)

	Site A			Site B		
	Local	Matched	Pooled	Local	Matched	Pooled
Total Reading	+	+	+	—		
Total Math	+			—	—	—
Spelling	+					
Language						
Raven's						
Coopersmith						
IARS (+)						
IARS (—)						
Word Knowledge	+	+	+			
Reading	+	+	+			
Math Concepts	+			—	—	—
Math Computations	+			—	—	—
Math Problem Solving						
Language Part A						
Language Part B						

Fifteen analyses were made for each Follow Through site reported--one for each variable listed in Table 1. An example of a research question might be stated as:

> Is the mean reading achievement test score of participating FT students greater than that of NFT students when the effects of:
>
> a. Fall kindergarten WRAT
> b. First language
> c. Family income
> d. Highest occupation in family
> e. Ethnic membership
> f. Sex
> g. Entry age
> h. Missing data code for WRAT
> i. Missing data code for income
> j. Missing data code for occupation
>
> are statistically controlled (where reading achievement is defined as the Total Reading score of the Metropolitan Achievement Test, which is comprised of the Word Knowledge and Reading subtests)?

As indicated in Table 1, comparisons were made between the FT students at each site and three different NFT groups: local, best-match, and pooled. Nine of these comparisons deal directly with the question of impact on reading: three each for Total Reading, Word Knowledge and Reading. The latter two of these are subtests that make up the Total Reading score. Of course, none of the six comparisons involving the Word Knowledge and Reading subtests are independent of the Total Reading Comparisons, but this is not specified in the table of effects or associated discussion.

When the specific research hypothesis addressed or the statistical hypothesis actually tested is not stated, the reader is left with the vague impression that everything that should have been controlled was controlled, and the numerous comparisons reported must have assessed the FT program effects on reading rather comprehensively. We are sure that the authors did not mean to leave that impression, and they assumed that any sophisticated reader would interpret their analyses and interpretations appropriately and with much caution--given the numerous caveats and explanations included in the first part of the report. But, in any 400 page report with an additional 400 pages of appendices, the reader will have difficulty figuring out how the scores that define reading were obtained, what they represent, and what the evaluators think they represent. The same problem exists for each of the ten or more covariates.

This is a complex and difficult problem faced by every evaluator at one time or another, and we do not want to address issues about the role of evaluation and of evaluation reports. Our concern is that evaluation reports describe as clearly as possible the evaluation activities undertaken to answer the general evaluation questions and communicate as precisely as possible the relationships between the general evaluation questions, the research hypotheses, and the statistical hypotheses actually tested. Ambiguity in a massive, complex evaluation tends to communicate to the reader that everything was done that could possibly be done and the evaluator's conclusions are the "best" interpretations, if not the only appropriate interpretations of the data. There are always pressures to make the evaluation as convincing as possible, whether positive or negative results are obtained, because the client paid for the evaluation. This often results in gross overstatements of findings or of the confidence one should have in the findings and often does not represent well the situation that is being evaluated. By specifying the general evaluation questions, the research and the statistical hypotheses,

and the evaluator's view of the relationships among them, both the strengths and weaknesses of a complex evaluation can be clarified. The limited empirical information presented in the resultant evaluation report can then be used more appropriately by decision makers and be more useful to educational professionals.

Are the Variables Defined, the Rationale and the Procedure for Selecting the Measures Described, and are the Relationships Among the Measures, Variables, and Evaluation Questions Specified?

In general, three relationships are of concern in the measurement area: (a) variable/domain, (b) instrument/variable, and (c) instrument/domain. Each of these has associated with it an inferential gap that must be bridged in order to relate the empirical results to the intended purposes of the evaluation. The rationales that delineate these relationships must be specified in the report so the reader can best assess the adequacy of the instrumentation.

A major issue in the measurement area is the conflicting considerations related to the "importance" and "scope" (Stufflebeam, Foley, Gephart, Guba, Hammond, Merriman, & Provus, 1971) of the data collected and reported. Importance deals with emphasizing the most important information in a particular situation and eliminating that which is not valued. Scope is the concern about the entire range, or comprehensiveness, of the information included in the evaluation. Decisions must be made about each possible type of evaluative information and datum to be included. As decisions are made about domains, variables, and measures, practical considerations of time, money, adequacy of measurement procedures, etc., tend to limit the evaluation to the most important variables and measures. At the same time, concerns about adequately fulfilling the purposes of the evaluation

tend to expand its scope. But, numerous modifications and compromises in each area are always made.

In the national FT evaluation, two domains (cognitive and noncognitive) were identified for student outcomes.[3] Some of the variables and measures used to assess the domains are listed in Table 2.

Table 2
Domains, Variables, and Measures Used in National Evaluation of Follow Through Program[a]

Domain	Variable	Measure
Cognitive	Total Reading	Metropolitan Achievement Test
	Total Math	
Cognitive	Total Reading	Metropolitan Achievement Test
	Total Math	Metropolitan Achievement Test
	Spelling	Metropolitan Achievement Test
	Language	Metropolitan Achievement Test
	Problem Solving	Raven's Progressive Matrices
Noncognitive	Self-Concept, or Self-Esteem	Coopersmith
	Locus of Control	Individual Achievement Responsibility Scale

[a]These were used to assess third-grade affects in the Abt evaluation (Stebbins, 1976).

[3]SRI identified two domains (cognitive and noncognitive) as opposed to the three domains (basic skills, cognitive conceptual skills, and affective) identified by Abt. In this paper, to simplify the discussion we deal only with the cognitive/noncognitive distinction--even though the three domains more adequately match the different sponsors' objectives.

All measures used in an evaluation must be specified and described, and the specific variables constructed from these measures must be defined. The specification of the variables and the measures of them can usually be done easily by using a table such as Table 2. When the variables are defined as tests or subtests of standardized tests, a short description of the test and the scores actually analyzed is usually adequate to enable the reader to understand how each variable is being operationally defined.

Whenever an evaluation is planned, a wide range of domains and variables are initially identified for possible inclusion. Often domains, variables, and measures are excluded during the selection process. The evaluation contractor is usually most knowledgeable about the compromises and deletions that are made. A discussion of this selection process is seldom, if ever, included in an evaluation report. Thus, the best thinking about this problem and the rationales for the decisions are lost to the field and to society. They are also not available to the readers, including major decision makers in Congress, who need that information so that they can more appropriately assess the relative value and importance of the conclusions of an evaluation report as they relate to decision alternatives.

An example of such a discussion appeared in Design for the Individualized Instruction Study (Cooley & Leinhardt, 1975b). The first two pages of their rationale for excluding noncognitive variables in their evaluation design are included as Appendix B of this paper. It indicates the steps that were followed and the criteria they used to arrive at their recommendation. In Section 3 of their report, Cooley and Leinhardt present their rationale for using a standardized achievement test to assess cognitive outcomes. The criteria utilized to compare possible tests are delineated. The actual test reviews are included in an appendix of their report, where the subtests of each achievement

battery, the psychometric characteristics, the norms available, and other characteristics are described. However, there is very little discussion by the authors of the inadequacies of the test battery that was to be used in the evaluation.

This example gives a rationale for and describes the relevant relationships between the domains, variables, and measures to be included in this evaluation. In our view, it would have been helpful to indicate more fully the strengths and inadequacies of the achievement battery in assessing the specific variables and the cognitive, or achievement, domain.

Since neither Abt nor SRI described the procedures and rationales that led to delineating the variables and measures utilized in the FT evaluations, we have identified some aspects that seem to have been considered.

1. Follow Through is an attempt to extend the positive effects of the Head Start program. Variables similar to those investigated in the Head Start evaluation should be included.

2. Follow Through as a compensatory education program has as its primary emphasis the improvement of students' basic skills, which in the first three grades are reading and mathematics.

3. The 22 FT sponsors have discernably different approaches to early childhood education. Domains and variables were identified from their main program objectives.

4. Given the amount of time and money allocated for the FT evaluation, only the most important and usable variables could be investigated. Thus, some important variables that are of interest could not be included because valid and reliable measures of them were not available.

A discussion of each of the considerations used to make decisions and judgements on the adequacy of the domains and variables is needed

if the complex analyses and interpretations are to be meaningfully understood and utilized. The evaluation report should indicate how these and other considerations affected variable selection and should include the rationales for the choices made. When these considerations are not included in a large complex evaluation, the reader is often left with the impression that all important domains and variables were included in the evaluation, and the measures used did adequately (and comprehensively) represent them.

What Criteria Were Utilized to Decide if a Specific ANCOVA Analysis Should be Made and Interpreted?

1. To what extent does each comparison meet these criteria?

2. What are the effects on the interpretation of results of the failure to meet the criteria?

The primary reason that ANCOVA-type procedures are used in evaluation settings is to adjust for, or statistically control, other likely explanations for the outcomes that are assessed. But, alternative explanations are still present after the analyses have been completed, whether or not randomization was used. When there is little or no experimental control--such as in naturalistic field studies or evaluations--outcomes are even more difficult to interpret and explain.

In naturalistic field settings, such as FT, deviations from the assumptions and conditions necessary to apply and interpret ANCOVA with some degree of precision, such as homogeneity of regression or similarity of groups in the analysis, are often present. The evaluator must attempt to assess the extent of the deviations and to delineate criteria for deciding when a specific analysis should not be interpreted. Establishing the specific values for criteria is an admittedly subjective

process, as there is little guidance available in the literature.[4] These values must be based on the purposes of the evaluation and on the specific situations in which the evaluation is occurring. In this section, we discuss several criteria that should be considered, and present methods of reporting them.

The first consideration that must be made, especially in a longitudinal evaluation, is whether the data in hand are representative of the situation being evaluated. In the FT evaluation, non-random attrition was often a major problem. After three to four years, less than 25 percent of the initial FT sample had complete data in some sites. A criterion that was implemented by the Abt evaluation team was that both the FT and NFT groups be comprised of at least 12 students. This small sample size would, of course, overfit the statistical model, especially when seven to ten covariates were used; but, at least the criterion value (12) was explicitly stated. The actual sample sizes of the samples included in specific analyses can usually be presented in a table summarizing the results (or "effects," in Abt's terminology). In addition to sample size, this table should report what proportion of each site's participating students comprise its sample. The size of this proportion directly influences the appropriateness of conclusions drawn from the analyses. The question of proportion of participants needed is a related but complex concern that will not be addressed here, but should be considered in each evaluation setting. Other considerations about the representativeness of the data that might be of concern are discussed in the final section of this paper.

---

[4] Implicit in much of the literature on ANCOVA is that it should not be used in non-experimental situations, but this is extreme. Selective application and cautious interpretation are a more practical and useful approach to using this and other statistical methods.

A second set of criteria are the assumptions of ANCOVA. The four considerations that are usually important were identified in Abt's FT report:

1. The covariates are uninfluenced by treatment;
2. The distribution of the covariates is not grossly different across groups;
3. The relationships between covariates and criteria are the same (homogeneous); and,
4. The covariates are perfectly reliable. (Stebbins, 1976, p. A-58)

Each of these assumptions was investigated or discussed in Abt's evaluation report. The first assumption was investigated by rerunning ANCOVA comparisons without the one covariate (WRAT) that they felt could be influenced by the treatment. Violation of this assumption meant that the portion of the treatment effect that was confounded with WRAT was being inappropriately removed. The report states:

> If the WRAT is influenced by the first few weeks of treatment, one might expect pretest adjustments to handicap the FT children. To test this we removed the WRAT from the covariate set and reran the local analyses. The results of these "no-WRAT covariate" analyses do not differ in any important ways from analyses which included WRAT as a covariate. We conclude from this comparison that the WRAT is probably not hindering our analyses of program effects. (Stebbins, 1976, p. A-59)

The rerunning of the ANCOVA analyses for each site was useful in addressing whether the use of the available WRAT data as a covariate affected the results obtained and conclusions reported. It is important to note that dropping any one of ten interrelated covariates is unlikely to affect the results of an analysis, given the relatively high correlations among covariates. Dropping the WRAT does not directly address the assumption that the covariates were not influenced by treatment. This assumption could be attended more directly

by giving the pretest earlier, such as in the previous year, before the program was implemented, or in the first two weeks of program implementation. Or, it could be addressed in a pilot study before the evaluation is undertaken.

Our own experiences at the Learning Research and Development Center (LRDC) have convinced us that much learning, as assessed by paper and pencil achievement tests, occurs in the first few weeks of our program. In a study in Pittsburgh area schools (Eichelberger, DiCostanzo, & Evaluation Staff, 1975), students using the LRDC curricula similar to that used in FT were assessed in the sixth week of school (as were a group of similar students) using the Metropolitan Readiness Test (MRT). The results obtained are reported in Table 3.

Table 3

Fall Metropolitan Readiness Test Results for Kindergarten Students in LRDC and Comparison Schools

School	Fall Mean	N
LRDC School 1	36.22	59
Comparison School 1	28.32	63
LRDC School 2	37.19	42
Comparison School 2	22.69	42
LRDC School 3	29.32	87
Comparison School 3	23.03	130

These results indicate that the three schools using the LRDC curricula during the first six weeks of kindergarten scored much higher than similar students who had not used that curricula. These results

suggest problems with the assumption that the fall kindergarten WRAT scores were unaffected by six weeks of treatment.

If the pretest was differentially affected by the treatments in FT, use of the ANCOVA-like procedures to test other assumptions and to adjust FT and NFT group differences in that evaluation might result in inappropriate conclusions. Elashoff (1969) suggested that analysis of variance be conducted on the covariate to test the assumption, but in the situation where testing occurs after four to six weeks of school, that procedure does not directly address the issue of the comparability of the groups prior to treatment. The implications of the failure to meet the assumption have not been well delineated at this time and deserve more careful consideration by evaluators using this technique.

If the evaluator's concern is to assess the effect of using a specific covariate (such as the WRAT) on the results obtained, then rerunning the analyses (without the WRAT) is useful. Whenever the "no-WRAT covariate" analyses result in changes in conclusions for a specific comparison, that fact and the associated results should be reported. The two sets of ANCOVA analyses might be performed at some specified level of significance (such as .05) and presented in a way that would reflect the different results that were obtained. When a large number of analyses and reanalyses are made, it is, of course, important to note the number of differences found significant as a proportion of the total number of comparisons made within each program or site.

The second assumption we will discuss--that the covariates were measured without error--has been theoretically studied, but what is known has seldom been applied in evaluation studies. Conclusions about educational programs drawn from empirical data may not represent the situation because of failure to meet this assumption. Evaluators often attempt to assess sampling error in a specific study, which is

often estimated from repeated use of the same measurement procedures. The conclusions may also be misleading about specific variables, such as academic achievement, because the measures inadequately assess important aspects of the variables. Neither of these problems can be solved with great confidence in an applied setting, so the complex adjustments are not attempted and the inadequacies in measuring the variables are overlooked.

There is also a tendency to overlook what Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, and York (1966) called measurement error. Measurement error " . . . includes such errors, among others, as ambiguities in definitions and in the questionnaire, failure to obtain required information from respondents, obtaining inconsistent information, mistakes in clerical coding and editing, errors occurring during the machine processing operation, and tabulation errors" (p. 561). In other words, it cannot be assumed that demographic and other "concrete" descriptive data are measured without error.

In Abt's FT evaluation, the authors indicated that "Variables such as sex, ethnicity, income, occupation, education, language, and age are all measured with a minimum of error. It is only the pretest which poses a problem" (Stebbins, 1976, p. A-60). With problems that exist in most self-report data--especially about variables like income and occupation among the low SES group--it is important that estimates of the reliability of these data be obtained and reported when they are used as covariates (see Elashoff, 1969; Lord, 1962).

An illustration of the difficulties that often arise in measuring what seem to be absolute entities occurred in a study at LRDC. The size of each classroom utilizing the LRDC program was to be measured

by making a sketch of the classroom area with the dimensions specified. Usually, this was done only once, because we felt we could reasonably assume that it was measured with a minimum of error. On one occasion, the measurement of the classrooms was asked for again in the same year. The results were not at all consistent, with shapes as well as dimensions changing. This experience has made us extremely cautious about the accuracy of all types of data--regardless of their presumed simplicity.

Coleman and his colleagues (1966), in the "Equality of Educational Opportunity" study, empirically investigated the systematic measurement error that resulted from selected parts of their procedures. Evaluators of all major longitudinal studies should consider estimating and reporting the measurement error associated with their data.

In Abt's FT evaluation, the degree of error in the WRAT pretest was investigated. The report stated that: "The reliability of the pretest was calculated by each Follow Through Sponsor-level sample by a measure of internal consistency (coefficient alpha) and is on the order of .90 across these samples" (Stebbins, 1976, p. A-60).

It is, of course, important to report the specific values for each group on which the analysis is run, because when the value is low for a specific group, the conclusions drawn from the analyses must be interpreted with even more caution. Even though 90% of the groups have very high reliability, specific sites may fall within a large range of values. The reader needs to know these values for specific analyses. It would be helpful if the evaluator initially set a reliability level (such as .80) below which the covariate would not be used (see Glass, Peckham, & Sanders, 1972, for reviews of studies that investigated this concern). This does not preclude a later decision to include a covariate that does not meet the criterion value, if there are unique

compelling reasons to do so. The reliabilities and associated cautions should be reported, or at least noted, in the text where the conclusions that they affect are reported.

A related point that must be raised at this time is the questionable use of coefficient alpha as an estimate of error in a covariate, such as the WRAT. This is not intended as criticism of Abt's use of it, given the type of data available to them and their general situation. In fact, Abt's attempts to investigate the adequacy of the FT data for the ANCOVA model are to be commended. But, it is important that better methods be identified and utilized for testing the assumptions and setting appropriate criteria. Tukey (1954) and Wold (1956) explicated problems that arise as data analysis moves from experimental to observational data, of which every researcher must be aware. Additional work on these problems is needed for clarifying their implications for decision-oriented research.

As previously indicated, no criteria related to the first and fourth assumptions were specified and used in the national FT evaluation, although both assumptions were addressed. The three criteria that the Abt evaluation used and reported to indicate that "the adjustment produced by ANCOVA may be misleading" (Stebbins, 1976, p. A-71) were:

1. when the relationship between a given covariate and outcome is different for the treatment and comparison groups being analyzed;
2. when the pretest difference between the treatment and comparison groups being analyzed is greater than five points (about one-half of a standard deviation); and
3. when the percent of those attending preschool in each group differed by more than 50 percent. (p. A-72)

The first criterion is essentially the ANCOVA assumption that the treatment groups have a common regression surface. The second is an indication that the treatment groups were not drawn from populations with the same covariate distributions. The third criterion could also be viewed as questioning the assumption that the groups were initially similar (see Campbell & Erlebacher, 1970, for how this can lead to erroneous conclusions). When any of these conditions existed for a comparison, or set of comparisons, their corresponding results were "greyed-out" in the effects table. The specific criterion values that were used in greying-out a particular comparison are presented in the text of the report, and which of these were violated in an analysis is specified in the effects tables. This is vastly superior to presenting all information in a table with no indication that some of the results are questionable. In fact, in certain situations it may be more appropriate to report only the results of analyses that do meet all of the minimum criteria, rather than merely greying-out certain results, since a reader tends to assume that all data reported are meaningful.

Note that these three criteria have associated with them explicit values or decision rules (such as statistical significance). Although the reader may disagree with the specific values set by the evaluator, s/he knows when the evaluator thinks the results are interpretable and what the specific criteria are on which the decisions are based. Testing for the violation of the conditions and assumptions needed for meaningful interpretation of ANCOVA results should be done in an experimental setting; however, such testing is imperative in an evaluation or naturalistic setting, where naturally confounded variables are almost certain to be present, and little control of the situation is possible.

Considerations should not be limited, however, to those discussed and dealt with in this paper, or in the Abt report. Others that may be relevant in your setting may have been excluded in the FT evaluation.

For example, can the criterion-covariate regression be expressed in linear form for each treatment group? If this condition is violated, i.e., the data cannot be transformed to linear form, comparison of two estimated treatment means will be biased. Elashoff (1969) notes "that the effect of nonlinearity is most severe when random assignment to groups is not possible or protection against non-normality in the y's is lowest" (pp. 390-391). This condition can be tested by examining increases in explained variation when higher order terms are included in the regression equation. Again, the evaluator should specify the exact criterion value for the test.

In summary, the evaluator of any program working in a naturalistic setting will find that the data available will deviate from assumptions and conditions necessary for direct interpretation of ANCOVA results. By recognizing this fact before implementing the evaluation, detailed guidelines can be designed based on the purpose of the evaluation and the evaluation setting. As more data and knowledge are gained from a program--especially a longitudinal one--modification in these criteria may be necessary. But, these guidelines with their associated rationales, their subsequent changes, and their implications for the conclusions drawn are needed by the reader to understand and interpret the results of an evaluation.

What Criteria Were Utilized to Determine If a Covariate Was to Be Included in a Specific Analysis?

Too often covariates are included indiscriminately in a set of ANCOVA analyses without knowledge of the local conditions or a theory of how the variables interrelate (see Cooley & Lohnes, 1976, for a discussion of the importance of theory in evaluative research). This usually results in conservative estimates of treatment effects, due to confounding of the covariates with the treatment. We believe that the

selection of covariates for an analysis should be based on a logical rationale, preferably one that is a part of a broader theoretical framework. Presenting the logical process used to identify candidate covariates indicates that the evaluator has broadly conceptualized the evaluation problem. Also, unique conditions in a specific situation often require decisions to be made about the inclusion or exclusion of a covariate in a specific analysis. In addition to a theoretical basis for including covariates, guidelines for excluding them are needed. These guidelines for including or excluding covariates are usually based on the several assumptions of ANCOVA listed in the previous section.

All covariates that were assessed in the evaluation and considered for use in a specific analysis should be listed in the report. The rationales for their inclusion should also be presented. This point was discussed extensively under variable identification and measurement in section two of this paper.

In large complex evaluations, numerous covariates are assessed, but the specific ones used in different analyses often vary. This variation is usually due to failure to meet one of the ANCOVA assumptions, or to conditions unique to the situation, such as only one race involved. When the results of an ANCOVA analysis are presented, a list of the covariates considered for use and the reason for excluding any from the analysis should be reported. A hypothetical example of such a list is presented in Table 4.

The criteria to decide whether or not a covariate should be included in a comparison should not be limited to the assumptions of ANCOVA. For example, one criterion not utilized in the Abt report was the degree of relationship between the covariate and outcome variable. This is an important factor for assessing the effectiveness of the covariate.

Table 4

Hypothetical Table of Covariates Considered for an ANCOVA Analysis and Reasons for Dropping Those That Were Excluded

	Criteria Failed	Covariates Included
Fall Kindergarten WRAT	1[a], 4	
Preschool Experience	2	
Sex		X
Ethnic Membership		X
Occupation	4	

[a]The numbers refer to the assumptions of ANCOVA listed in section 3.

Cox (1957) compared the precision of blocking versus covariance for different values of rho (ρ), i.e., the correlation between the covariate and outcome variable. Cox concluded that if $\rho < .4$, blocking is preferable to covariance analysis; if $.6 < \rho < .8$ covariance is somewhat better; and if $\rho > .8$ covariance analysis is appreciably better. Although other factors affect this relationship (e.g., shape of covariate distributions), it is an important consideration that should be used as a criterion for judging potential covariates. This consideration is, of course, secondary to the purpose of the evaluation and the specific questions being addressed in any study.

The general information on choosing covariates that would appear in a table similar to Table 4 should be supplemented with the specific values of correlation coefficients for the selected covariates. This could be efficiently incorporated into a table exhibiting the information necessary for the reader to reconstruct the regression equations of the comparisons that were actually made. A pertinent illustration can be found in the appendices of the SRI report (Emrick et al., 1973). In Table 5, the correlation with the dependent variable, the raw regression

Table 5

Cohort I, Kindergarten, Regression on Control Variables: Pupil Outcomes for Project Analyses ($N = 330$, Residual $df = 258$)[a]

			Achievement Regression Coefficients				WRAT Regression Coefficients				Affect Regression Coefficients				Absence Regression Coefficients			
Covariable	Mean	S.D.	r_0	Raw	STD	S.E.	r_0	Raw	STD	S.E.	r_0	Raw	STD	S.E.	r_0	Raw	STD	S.E.
Fall 1969																		
Quant. Prescore	-.027	.519	.690	14.470	.383	3.033	.655	7.243	.373	1.638	.108	.594	.219	.318	.000	.269	.033	.992
Cog. Process Prescore	-.029	.608	.440	-4.669	-.120	2.255	.410	-2.688	-.135	1.218	-.021	-.516	-.186	.237	-.009	-.368	-.044	.738
Reading Prescore	-.029	.535	.699	10.019	.273	2.557	.670	5.238	.278	1.381	.063	-.228	-.087	.268	.011	.125	.015	.837
Language Prescore	-.032	.476	.648	10.001	.243	2.811	.616	5.006	.237	1.410	.075	.056	.018	.274	-.002	.065	.006	.864
Affect Prescore	-.047	.426	-.041	-4.143	-.089	1.835	-.083	-3.069	-.129	.991	.280	.964	.288	.193	.014	.203	.020	.600
Av. Pupil Age (Months)	84.14	1.78	.216	-.095	-.008	.468	.197	-.136	-.024	.253	.047	-.007	-.008	.049	-.017	-.036	-.016	.163
% Classroom Male	49.03	21.59	-.067	-.032	-.035	.036	-.062	-.013	-.028	.019	.037	.000	.016	.003	-.200	-.041	-.213	.011
% Classroom Black	74.19	15.82	-.064	-.046	-.037	.065	-.050	-.047	-.074	.046	.017	-.008	-.095	.008	-.090	.007	.027	.027
% English 1st Language	94.28	10.32	.006	-.094	-.049	.092	.003	-.054	-.056	.044	.010	.000	.001	.008	-.021	.007	.018	.026
% Preschool (or No. Mos.)	56.58	21.67	.008	-.090	-.099	.036	.001	-.048	-.105	.019	.105	.006	.077	.003	.033	.007	.036	.012
% Parents w/o HS Dipl.	69.92	19.28	-.215	-.019	-.019	.044	-.188	.001	.003	.024	-.107	-.004	-.056	.004	-.016	-.018	.081	.014
% Parents w Skill Occup.	35.36	19.63	.101	-.043	-.043	.042	.103	-.012	-.024	.022	.130	.008	.123	.004	-.026	-.009	-.042	.013
% Parents Black	75.29	15.20	.017	.051	-.039	.088	.048	.073	.110	.047	.057	.013	.148	.009	-.139	-.053	-.193	.028
% Parents Poverty Eligible	64.48	18.98	-.220	-.029	-.028	.045	-.178	.003	.006	.024	-.115	-.007	-.104	.004	.100	.014	.063	.014
% Head Household Employed	66.75	19.10	.254	.169	.164	.060	.236	.082	.155	.027	.026	-.005	-.072	.005	-.144	-.032	-.144	.016
% Head Household Male	58.68	19.84	.115	-.082	-.083	.048	.103	.036	-.071	.026	-.003	.002	.030	.005	-.063	-.001	-.008	.015
Summary Statistics																		
Mean				126.91				64.98				17.53				13.24		
Variance				383.56				101.22				1.97				18.00		
Multiple R				.778				.761				.390				.315		
R^2				.606				.564				.152				.100		
Variance with Cov's Eliminated				160.68				46.88				1.78				17.21		

[a] From Emrick et al. (1973, p. B-18).

weight, the standardized regression weight, and the standard error of the regression coefficient are listed for each covariate. This should be accompanied by information describing and elaborating on the ANCOVA comparison made, such as unadjusted and adjusted outcome variable means,[5] the standard error of the adjusted difference, the sample sizes, and the actual results of the comparison in the form of a computed statistic or confidence interval.

To summarize, we have recommended that the evaluator delineate a logical rationale for the selection of variables as candidate covariates, preferably based on an overall theoretical framework. Guidelines should be specified for deciding when covariates should not be included in an analysis. The criteria specified should include, but not be limited to, the assumptions of ANCOVA. The results of this decision process could be presented in a table similar to Table 4. Finally, the types of data that should be reported for each covariate and comparison that allow the reader to assess the interpretations derived from the ANCOVA comparisons have been discussed. In large complex evaluation reports, we have often found it very difficult to identify the variables that were included in an analysis, let alone find the reasons why a particular variable was or was not included. Criteria that might be used to decide which covariates to exclude from an analysis, and methods for clearly presenting the associated information in an evaluation report, need further investigation and development.

---

[5]Some indications that the adjusted means have no intrinsic value, and that comparison of the means with their associated unadjusted means is not usually meaningful, should be included in the report. Of course, the difference between the adjusted means is used in the computation of statistical significance.

Are the Different Groups Included in the ANCOVA Analyses and Their Educational Experiences Described Adequately in the Report?

Much of the technical information required to assess the appropriateness of the interpretation of the evaluation results has been specified in the preceding sections. We have commented on the need for the evaluator to: (a) link the data analyses and research hypotheses to the general evaluation questions, (b) specify and link the measures and variables utilized in an evaluation to the domains of interest, (c) state the criteria used to decide if an analysis should be made and interpreted, and (d) specify the criteria used to select covariates for a specific analysis. In addition to these concerns, several others pertaining to the groups' characteristics and experiences are needed for the conclusions to be interpreted appropriately.

Knowledge of the educational conditions and treatments that the different groups experienced is of central importance in interpreting the results of any program evaluation. A detailed description of the programs experienced by students is a major undertaking, as evidenced by the extensive work in FT of Stallings (1973) at SRI and of Cooley and Leinhardt (1975a) at the Learning Research and Development Center. Obviously, the extensiveness of the program descriptions represented by these studies usually cannot be achieved when conducting an impact evaluation, but the identification of some essential context and program variables should be made by evaluators in any setting. Ignoring differences between intended treatments and those actually experienced, and between characteristics of the "experimental" and "comparison" groups, can lead to erroneous conclusions about the relative impact of the variables being evaluated. Also, little or no knowledge is gained about how the obtained outcomes had been affected by important program variables.

Follow Through again provides a relevant example. The FT evaluation was intended to assess the impact of the FT program on participating children, as compared to the impact of "regular" school experiences that did not include innovative educational programs. However, the "regular" school programs serving the NFT comparison children often included other compensatory programs, such as Title I, which at times utilized educational materials and practices similar to those in some Sponsors' FT instructional models. As a result, when an FT/NFT comparison is made, the appropriate interpretation of the results is not immediately apparent. Differences between FT and NFT groups and their educational experiences must be integrated with the reporting of results. A hypothetical example might be, "The NFT children at the Oshkosh, Alaska, site were similar to the participating FT children at the site on all entry characteristics measured. Because the NFT children were from families whose incomes were very low, they qualified and participated in the Title I federal compensatory education program. This involved supplemental instruction in arithmetic and reading and additional aid. . . ." This type of information is needed by the reader to interpret the results with respect to the educational variables actually being assessed and the degree to which differences in outcomes might be expected.

In addition to considerations about the comparability of the educational conditions and materials the different groups experience, the evaluator must report information about the similarities or differences between the groups experiencing the program being evaluated and those comprising the comparison group. In previous sections, the necessity to report raw and adjusted means on the covariates and the dependent variables was noted. Suggestions of how and where to report the information was also indicated. Other aspects of each unique evaluation setting must also be taken into account. Within FT, some of these considerations relate to attrition and missing data, program requirements

for participation, and local implementation and utilization of the program.

Attrition and missing data commonly affect the final composition of the groups being compared. Attrition occurs when a participating student moves out of the FT classroom. Missing data occurs when one or more measurements for a participating student are missing. Due to these two factors, the composition of groups in the FT evaluation has been shown to undergo drastic changes during the course of a four-year educational program. For example, the Abt report states that "approximately 50 percent of the FT and NFT children who are tested in the kindergarten year of Cohort II were not present at the end of third grade" (Stebbins, 1976, p. A-47).

Empirical investigation can be utilized to determine whether attrition or missing data bias a comparison. The Abt evaluators compared rates for FT and NFT students at each site, using their pretest scores and family income data. Five sites were found for which attirtion significantly changed the difference between groups' pretest scores, and three sites were identified for which attrition altered the FT/NFT difference in mean income. No explanation was given in the report for the selection or limitation of the investigation to these two variables.

A procedure was used in the Abt report to estimate values for the missing data for covariates. Whether or not a covariate value was estimated was then noted in the analysis. Several advantages to this procedure were noted in the Abt report:

> It avoids the risk of nonrepresentativeness due to dropping children;
>
> it avoids the loss of statistical power due to reduced sample size;
>
> it uses the information contained in the absence-presence of the variable;

> and it uses the information present on other variables for children who might have been dropped otherwise. (Stebbins, 1976, p. A-51)

In any large-scale longitudinal evaluation, the evaluator will have the task of selecting from numerous alternative approaches for handling missing data, including dropping such persons from all analyses. Each situation will dictate considerations that will influence the decision rules for handling missing data. We suggest that these rules and their rationales be made explicit. How the estimation of missing data affects the assumption that the measures are perfectly reliable and how the interpretation of the results might be affected must be considered.

Federal requirements for the FT program also affected the composition of the FT and NFT groups:

> Children enrolled in early elementary grades may participate in [FT] projects. . . . At least 50 percent of the children in each entering class shall be children who have previously participated in a full-year Head Start or similar quality preschool program and who were low income at the time of enrollment in such preschool program. (Federal Register, 1975, pp. 11714-11715)

As a result, entering kindergarten children could not be randomly selected for participation in FT. At some sites, those students below poverty level were assigned to the FT classroom while students from higher income families were assigned to the regular classrooms and often become part of the NFT comparison groups at the site. The descriptive data do indicate the existence of this systematic bias caused by program requirements (see Table 6).

Local decisions about implementation and utilization of the FT program are more difficult to document, but no less a problem for adequate interpretation of results. For example, local administrators often used FT as a remedial program. Students who were

Table 6

Descriptive Characteristics of the National Population, the Follow Through Sample, and the Non-Follow Through Sample

	National	FT	NFT
Median Income	$9590	$4450	$6060
% Minorities	13%	86%	79%
% Preschool	9%	81%	57%
Fall WRAT	NA	29.7	29.4

repeating a grade or who had special needs were often placed in the FT classroom. The Abt report also indicated that a systematic bias exists against the FT group: "In most cases the Follow Through participants were selected from among the 'most difficult' in the community . . . some communities chose to include the mentally handicapped and/or emotionally disturbed" (Stebbins, 1976, pp. A-12 - A-13).

Although this was not the case at all sites, it does document that the "more difficult" students were placed in FT classes. The effects of these differences, such as the inclusion of emotionally disturbed children and grade repeaters, usually remain unknown because they are not assessed by the covariates and are not investigated in other ways either.

These examples emphasize the need for detailed descriptions of the groups being compared and their educational experiences. This information should be coordinated with the reporting of results at the site level, since program interpretations depend upon the similarity of the groups. The results section of the Abt report did describe FT/NFT group comparability both in tabular and prose forms. For

example, the FT and NFT groups at a particular site were described in the results section for that site:

> The FT group is also well below sponsor average in income, while the NFT is about average for this sponsor. . . . The two groups are a fairly close match on entry WRAT and ethnic composition, though the NFT income level is considerably higher than the FT level. (Stebbins, 1976, p. A-195)

This information permits the reader of the report to make more appropriate interpretations of conclusions and other summary statements made by the evaluator about a specific site by making him/her aware of the similarities in entry WRAT and ethnic composition and the considerable difference in income.

Summary

The purpose of this paper was to indicate some specific information that should be included in an evaluation report when ANCOVA-type techniques are used in order to allow the reader to assess the adequacy of the analyses and the appropriateness of the evaluator's interpretations of the results. A recurrent theme of this paper has been that the evaluator must recognize that the application of ANCOVA in evaluation settings requires a more elaborate analysis and reporting strategy than in experimental studies due to the failure to meet assumptions of ANCOVA precisely and the existence of numerous plausible alternative interpretations of the results. The evaluator must recognize that important aspects of the evaluation should be described in detail, i.e., setting, treatments, characteristics of the participants and nonparticipants, and their educational experiences.

The major points elaborated in the paper are summarized below. Those activities that evaluators often fail to carry out, or criteria that are sometimes not specified in the section of the report where they could be mose useful, are emphasized:

1. Specify the hypothesis actually tested by an analysis rather than only relating the analysis to a general evaluation question.

2. Describe the variables used in each analysis, the rationale and procedure for selecting the measures, and the relationships among the measures, variables, and evaluation questions.

3. Use explicit criteria to decide whether or not to make a specific analysis and report the extent to which an analysis meets those criteria.

4. Use explicit criteria to decide whether to include a covariate in a specific analysis.

5. Describe the groups included in the ANCOVA-type analysis by reporting:

   (a) adjusted and unadjusted raw or standard means on the dependent variable(s) for each group,
   (b) summary statistics for each group on the covariates used in each analysis, and
   (c) a detailed description of the educational experiences of the program groups and of any comparison groups.

These points were made in an effort to reduce the ambiguity that often ensues when reporting the results of ANCOVA techniques in complex longitudinal evaluations. As indicated by Tukey (1954), "Experimental statisticians should be honest and expository about the relation of precise assumptions and exactly optimum solutions to real situations" (p. 719). These considerations are intended to improve evaluators' abilities to communicate their findings accurately to the nonstatistically oriented reader. A special effort is needed to indicate the limitations of an evaluation as well as its strengths so that a more balanced and accurate picture of a program and its effects is presented to the decision maker, who may be puzzled and awed by the mathematical procedures.

References

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

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

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

Cooley, W. W. Explanatory observational studies. Educational Researcher, 1978, 7(9), 9-15.

Cooley, W. W., & Leinhardt, G. The application of a model for investigating classroom processes. Pittsburgh: University of Pittsburgh, Learning Research and Development Center, 1975. (a) (LRDC Publication 1975/24).

Cooley, W. W., & Leinhardt, G. Design for the individualized instruction study: Final report. Pittsburgh: University of Pittsburgh, Learning Research and Development Center, 1975. (b) (NIE Contract No. 400-75-0071)

Cooley, W. W., & Lohnes, P. R. Evaluation research in education. New York: Irvington Publishers, 1976.

Cox, D. R. The use of a concomitant variable in selecting an experimental design. Biometrika, 1957, 44, 150-158.

Egbert, R. L. Planned variation in Follow Through. Paper prepared for the Brookings Panel on Social Experimentation, Washington, D. C., 1973.

Eichelberger, R. T., DiCostanzo, J. L., & Evaluation Staff. Evaluation of the LRDC individualized programs in the Pittsburgh Public Schools. Unpublished manuscript, University of Pittsburgh, Learning Research and Development Center, 1975.

Elashoff, J. D. Analysis of covariance: A delicate instrument. American Educational Research Journal, 1969, 6(3), 383-401.

Emrick, J. A., Sorensen, P. H., & Stearns, M. S. Interim evaluation of the national Follow Through Program 1969-1971: A technical report. Menlo Park, Cal.: Stanford Research Institute, 1973.

Federal Register. Follow Through Program. Washington, D. C.: U. S. Government Printing Office, April 21, 1975 (Part II).

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

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

Johnson, L. B. Special message to the Congress recommending a 12-point program for America's children and youth. Public Papers of the Presidents, February 1967, 150-160.

Lord, F. M. Elementary models for measuring change. In C. Harris (Ed.), Problems in measuring change. Madison, Wisc.: University of Wisconsin Press, 1962.

Sorensen, P. H., & Madow, W. G. A proposal for research: Longitudinal evaluation of the national Follow Through Program, 1969-70. Menlo Park, Cal.: Stanford Research Institute, 1969.

Stallings, J. A. Follow Through Program classroom observation evaluation 1971-72. Menlo Park, Cal.: Stanford Research Institute, 1973.

Stebbins, L. B. (Ed.). Education as experimentation: A planned variation model (Vol. 3). Cambridge, Mass.: Abt Associates, Inc., 1973. (USOE Contract No. 300-75-0134)

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

Tukey, J. Unsolved problems of experimental statistics. Journal of the American Statistical Association, 1954, 49, 706-731.

Wold, H. Causal inference from observational data: A review of ends and means. Journal of the Royal Statistical Society, Series A, 1956, 119, 351-390. Reprinted in M. W. Wittrock & D. E. Wiley (Eds.), The evaluation of instruction: Issues and problems. New York: Holt, Rinehart and Winston, 1970.

Wolff, M., & Stein, A. Six months later: A comparison of children who had Head Start, Summer 1965, with their classmates in kindergarten. New York: Yeshiva University, Ferkhauf Graduate School of Education, 1966.

APPENDIX A

(Quoted from Stebbins, 1976, p. A-46)

Given the need to provide information to decision makers, the essential problem becomes the development of an evaluation approach that will provide the most valid and comprehensive information possible. To this end, Follow Through evaluation planners (USOE, SRI) adopted a quasi-experimental design, selecting at each site a comparison group as similar to the treatment group as possible. Since this design does not suggest a single "appropriate" analysis, we have subjected the data to a variety of "approximately appropriate" analytic procedures, so as not to be overly confined by the drawbacks of the design. The multiple strategies approach anticipated the common and valuable practice of performing secondary analyses such as those performed on the Equality of Educational Opportunity data. Any single analytic treatment of quasi-experimental data is inevitably subject to well-founded methodological criticism, especially when the data are being used to assess the impact of major educational programs. Subsequent reanalyses using other techniques and approaches help to assess the validity of the original results. Usually, after several reanalyses have been accomplished and a body of literature accumulated, all available information is integrated to refine and clarify understanding of the problem (or program). Our analytic cross-validation anticipates some of the more obvious reanalyses and should provide other researchers with a broader basis for designing further thoughtful approaches to the Follow Through data.

APPENDIX B

(Quoted from Cooley & Leinhardt, 1975b)

Although the RFP calls for consideration of the "nonachievement factors which contribute to classroom environment," it does not spell out what these factors might be. It was suggested that the designer review this area and propose what definitions and instrumentation, if any, should be included in the Individualized Instruction Study. Our approach to this task has been two-pronged: (1) to determine whether non-cognitive student outcomes can and should be measured, and, (2) to determine whether it is possible and desirable to assess the effect of programs on the total classroom environment.

We do not recommend that non-cognitive student outcomes be assessed in the study for two reasons. First, although schooling, individualized or not, may indeed have an effect on some non-cognitive outcomes, the theoretical basis for such a belief is not well developed. Without a sound basis, it is futile to attempt to measure non-cognitive or social outcomes since it is not clear what to measure or how to make causal arguments if effects are found. A second argument against the test of social outcomes is that their measurement in the primary grades is still in a primitive state.

Our consideration of non-cognitive or social outcomes began with the generation of a list of outcomes that designers of instructional programs have claimed will be affected by their programs (e.g., self-concept, inquiry skills, autonomy). The next step was to locate instruments that purport to measure these specific outcomes. The short duration of the study ruled out the possibility of developing such instruments from scratch. Existing instruments were located, screened, and eliminated from further consideration if they failed to meet any one of the following criteria:

1. The instrument could not be highly correlated with reading and mathematics ability. If it were, it would measure little not already measured by the achievement test battery.

2. The instrument had to measure the social variables in question, i.e., it had to be valid as measured by standard measures of validity.

3. The instrument had to be reliable as measured by standard measures of reliability.

4. The instrument must have been designed or adapted for use in the primary grades.

5. The instrument must be usable from an administrative standpoint. This criterion would rule out instruments that are described in the literature but are otherwise untraceable, those that require an exorbitant amount of pupil/examiner time (in excess of three hours per pupil), and those that require a highly trained examiner or coder. A number of projective tests like doll-play were eliminated under this criterion.

The results of the search for an instrument that would meet these criteria were disappointing. Not one instrument of the many considered was totally acceptable. Table 3.1 lists some of the tests that were rejected and a criterion they failed. They may have failed other criteria, but this information was not recorded because the test reviewers eliminated an instrument upon failure to meet one criterion.